Universiteit Leiden
The Netherlands

Novel algorithms for protein sequence analysis

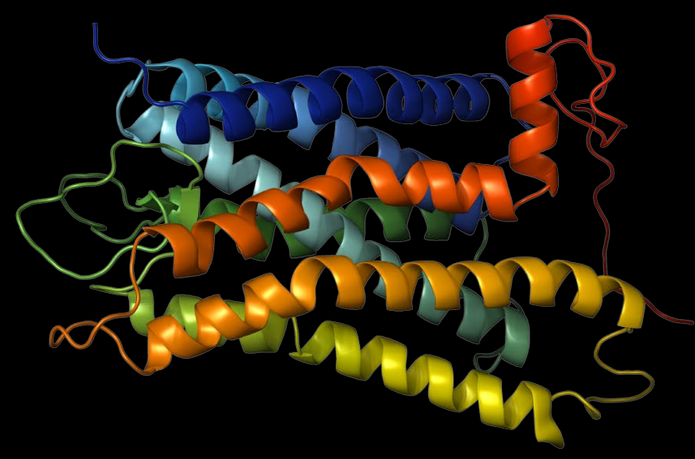
Ye, K.

Citation
Ye, K. (2008, December 18). *Novel algorithms for protein sequence analysis*. Retrieved from https://hdl.handle.net/1887/13355

Version:	Corrected Publisher’s Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/13355

Note: To cite this publication please use the final published version (if applicable).

Novel Algorithms for Protein Sequence Analysis



Kai Ye

Novel algorithms for protein sequence analysis

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 18 december 2008
klokke 13.45 uur

door
Kai Ye
geboren in HuBei, P.R. China
in 1977

Promotiecommissie

Promotor:	Prof. dr. A. P. IJzerman
Co-Promotor:	Dr. W. Kosters
Referent:	Prof. dr. G. Vriend (Radboud Universiteit)
Overige leden:	Prof. dr. M. Danhof Prof. dr. J. Kok Prof. dr. T. Hankemeier

The research described in this thesis was performed at the Division of Medicinal Chemistry of the Leiden/Amsterdam Center for Drug Research, Leiden University, Leiden, the Netherlands and financed by Leiden University, NOW (Horizon Breakthrough project) and the Dutch Top Institute Pharma (D1-105).

Contents

Chapter 1	General Introduction	5
Chapter 2	A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors	15
Chapter 3	Tracing evolutionary pressure	43
Chapter 4	Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting	59
Chapter 5	An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences	77
Chapter 6	Alignment independent phylogeny reconstruction – a cheminformatics approach	93
Chapter 7	Conclusions and perspectives	107
	Summary	113
	Samenvatting	116
	List of publications	119
	Curriculum Vitae	121
	Acknowledgements	122

Chapter 1

General introduction

Proteins are functionally important.

The major roles of proteins include transmitting signals (**receptor**) or materials (**transporter**) between compartments of cells and catalyzing chemical reactions (**enzyme**). Receptors are membrane (plasma or nuclear) bound and recognize a particular signal (light, a chemical compound, a peptide or even a protein) from one side of the membrane and invoke a biochemical response at the other side. They often have a binding site exposed on the cell surface and an effector domain within the cell, which may have enzymatic activity or may undergo a conformational change detected by other proteins within the cell. In this way, cells communicate with each other and hence creatures sense the environment and adjust accordingly to survive. For example, we see things because rhodopsin molecules in our eyes capture light of various wavelengths and we may smell and taste because of our olfactory and taste receptors. The Nobel Prize in Physiology or Medicine for 2004 was awarded to Richard Axel and Linda B. Buck for their discoveries of olfactory receptors (Malnic et al., 2004; Ngai et al., 1993).

Many ligand transport proteins recognize particular small molecules and shuttle them across the membrane to other locations of the cells. Coordinated material transport is essential for chemical reactions and signal transduction. When we drink water, we distribute it all over our bodies. The water channel proteins allow water to enter the cells. Ion channels, on the other hand, regulate the flow of ions across the membrane in all cells to establish and control the small voltage gradient across the plasma membrane of all living cells. Ion channels are especially prominent components of the nervous system since they conduct nerve impulse. Most toxins that organisms developed interact with ion channel pores to shut down the nervous system of predators or prey. In the search for new drugs, ion channels are also a favorite target. The Nobel Prize in Chemistry for 2003 was awarded to Peter Agre and Roderick MacKinnon for their discoveries of the water and ion channels of cells (Gouaux and Mackinnon, 2005; Kozono et al., 2003).

The best known role of proteins inside the cell is their duty as enzymes, which catalyze chemical reactions. Enzymes carry out most of the reactions involved in metabolism and catabolism, as well as in DNA replication, DNA repair, and RNA synthesis. Organisms transform incoming materials to produce energy and building blocks for themselves. Enzymes accelerate such chemical reactions which take too much time under body temperature in uncatalyzed conditions. The rate acceleration is often amazing. For example, orotidine 5'-phosphate decarboxylase, an extremely proficient enzyme, enhances the rate of reaction by a

factor of 10^{17} (78 million years without the enzyme, 18 milliseconds with the enzyme) (Radzicka and Wolfenden, 1995). The Nobel Prize has been awarded to enzyme researchers on many occasions. For example, the Nobel Prize in Chemistry for 1972 was awarded to Christian B. Anfinsen, Stanford Moore and William H. Stein for their contribution to the understanding of the connection between the catalytic activity of ribonuclease and its sequence and structure (Cuatrecasas et al., 1968; Moore and Stein, 1973).

Proteins are linear combination of amino acids

Proteins are linear polymers built from the 20 different L-α-amino acids. These amino acids possess common structural features, including an α carbon to which an amino group, a carboxyl group, and a variable side chain are attached. Only proline differs from this basic structure, as it contains an unusual ring to the N-end amine group, which forces the CO-NH amide moiety into a fixed conformation. The side chains of amino acids confer divergent biochemical properties which provide the driving force for protein folding and constitute a micro-environment on the protein structure for recognition of small molecules and the catalysis of chemical reactions. The amino acids in a polypeptide chain are linked by peptide bonds formed in a dehydration reaction. The sequential order of amino acids on the polypeptide chain is specified by the general genetic code. Once linked in the protein chain, an individual amino acid is called a *residue*, and the linked series of carbon, nitrogen, and oxygen atoms are known as the *main chain* or *protein backbone*. Due to the chemical structure of the individual amino acids, the protein chain has directionality. The end of the protein with a free carboxyl group is known as the C-terminus or carboxyl terminus, whereas the end with a free amino group is known as the N-terminus or amino terminus.

Since proteins are linear polymers of amino acids with directionality, we may use a one letter annotation to indicate each residue and a sequential order of these letters, or simply “sequence” for a protein.

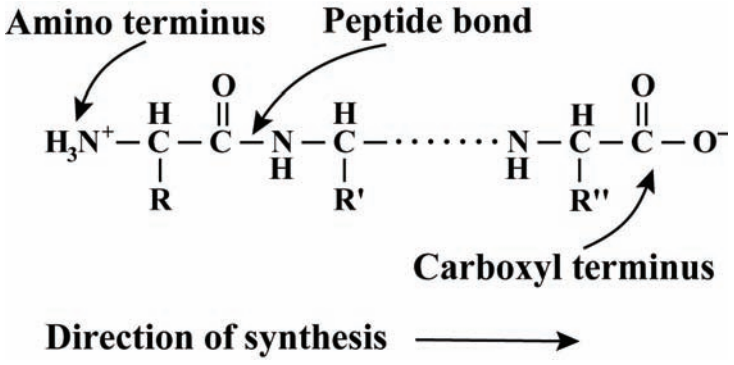


Figure 1. Proteins are linear combinations of amino acids with direction.

Protein family

Proteins do not emerge suddenly but evolve gradually from their ancestor proteins. We call proteins that evolve from a common ancestor a protein family. It is generally believed that proteins in the same family have similar biological function and three-dimensional structures. For example, the family of G protein-coupled receptors (GPCRs) is one of the biggest in our body. They are all membrane bound and have seven transmembrane α-helical domains, sensing molecules outside the cell and activate intracellular signal transduction pathways. Only two GPCRs, bovine rhodopsin and the β_2-adrenergic receptor, have their structures solved. Although the sequence (amino acid) identity between bovine rhodopsin and β_2-adrenergic receptor is as low as 11.0%, their structures are remarkably similar.

Protein sequence alignment

Each protein is characterized by its sequence. To predict the biological functions of one protein and the roles of its residues, we usually compare the sequence of this protein with similar protein sequences whose functions have been examined experimentally. For example, if a newly discovered protein is very similar to the human A_{2B} adenosine receptor, a G protein-coupled receptor, it probably also belongs to the G protein-coupled receptor family and a first experiment would be to examine whether it recognizes adenosine. Traditionally, to build links between residues among these sequences, sequence alignment is often used. In bioinformatics, a protein sequence alignment is the procedure of comparing two (pair-wise) or more (multiple sequence alignment) sequences by searching for a series of individual residues or residue combinations that are in the same order in the sequences. Each sequence is presented as a row across a page. The aligning process is a way of arranging the sequences such that similar or identical residues are placed in the same column, which will be called a position in this thesis. Non-identical or dissimilar residues are either placed in the same column as a mismatch, or opposite a gap in certain sequences. The first principle in the alignment process is to maximize the number of positions that have identical or similar residues. Mismatches can be interpreted as point mutations while gaps as insertions or deletions. A rough estimate of similarity between two sequences is the ratio of identical residues over the entire sequence in the alignment including gaps (sequence identity).

One may manually align two sequences but sophisticated algorithms are demanded for the alignment of dozens or even hundreds of sequences automatically. ClustalW was introduced to biologists in 1994 (Thompson et al., 1994). It quickly became the method of choice for its sensitivity and efficiency. ClustalW uses a predefined substitution matrix to assess the cost of matching two residues so that the score of matching depends only on the considered positions or their immediate surroundings. T-Coffee, another algorithm, is built on consistency-based schemes (Notredame et al., 2000). It first creates a collection of pairwise local and global alignments and then use this collection as a position-specific substitution matrix during a

regular progressive alignment. The final goal is to derive a multiple sequence alignment as consistent as possible with the collection (Notredame et al., 2000).

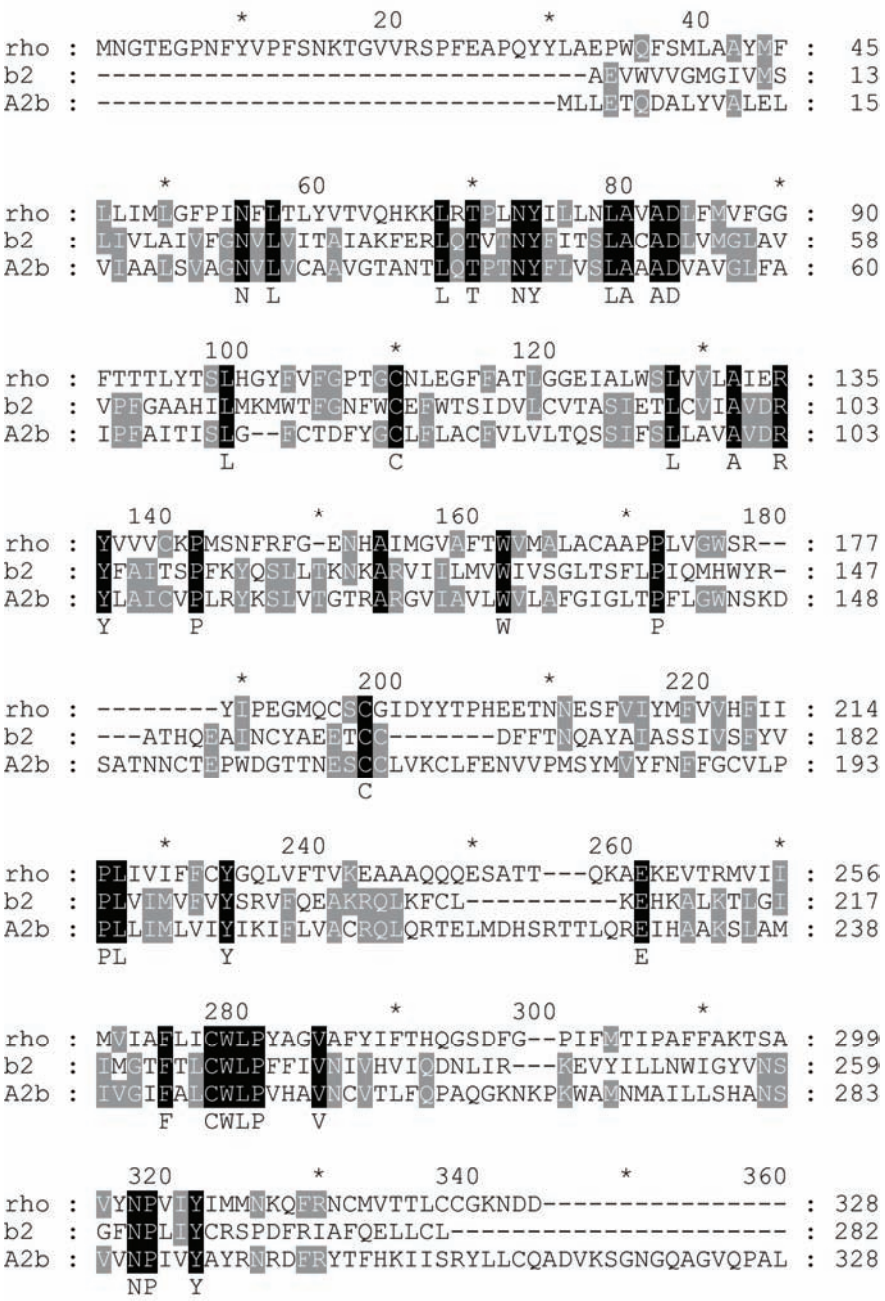


Figure 2 An example of a multiple sequence alignment. The sequences of bovine rhodopsin, β_2-adrenergic receptor and adenosine A_{2B} receptor were aligned with ClustalW (Thompson et al., 1994) and visualized in GeneDoc (http://www.nrbsc.org/gfx/genedoc/index.html).

An example of a multiple sequence alignment is given in Figure 2. The sequences of bovine rhodopsin, the β_2-adrenergic receptor and the adenosine A_{2B} receptor were aligned with ClustalW (Thompson et al., 1994) and after that the alignment was visualized in GeneDoc

(http://www.nrbsc.org/gfx/genedoc/index.html). The residues shared by all sequences are marked in black and are called conserved in this thesis. The combination of several conserved residues is called a pattern or motif in this thesis. For example, at the end of the second block in the example alignment, there is a short pattern LAxAD in which x indicates any of the 20 amino acids.

Phylogenetic analysis and alignment

Once a multiple sequence alignment has been built, the number or the types of residue substitutions may be used to illustrate the evolutionary history of these proteins. A phylogenetic analysis of a group of related protein sequences is a determination of how this group might have been derived during evolution. The degree of similarity between sequences is qualitatively related to the distance of one from the other in the phylogenetic tree. Roughly speaking, high sequence identity suggests that the two sequences in comparison started to differentiate late in evolution while low identity suggests that the divergence is more ancient. A simple procedure to reconstruct a phylogenetic tree first requires calculation of distances among the sequences in the multiple sequence alignment. Then one can recursively link pairs of sequences or sequence groups that have the shortest distance and then update distances between the new group and other sequences or groups. Such a tree construction algorithm is called Neighbor-Joining (Saitou and Nei, 1987).

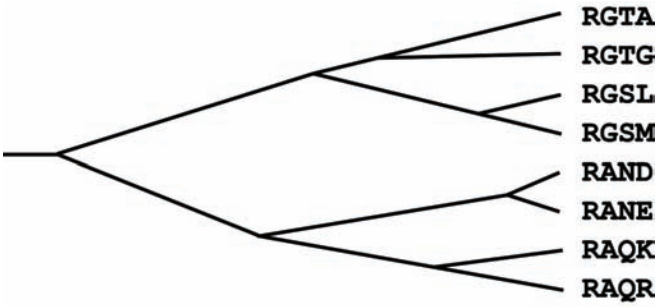


Figure 3 a hypothetic multiple sequence alignment and its correspondence phylogenetic tree.

PHYLIP is one of the most popular phylogeny reconstruction packages. It is available for free at http://evolution.genetics.washington.edu/phylip/programs.html. However the programs in PHYLIP can only be executed from a command line. MEGA3, an implementation of popular sequence alignment and phylogeny reconstruction methods with a convenient graphic interface, recently gained popularity in the biologists' community. We will briefly discuss designing novel algorithms versus a better implementation of established methods in Chapter 7 where we consider present conclusions and perspectives.

Motif, conserved site and specificity site

Starting from a multiple sequence alignment, one may not only construct a phylogenetic tree but also identify short fragments that show little variation among proteins in the alignment. For example, in Figure 4, there are 16 protein sequences from one protein family grouped into 4 clusters based on a biochemical property. There is always an arginine residue at position a. Thus position a is conserved and probably important for the overall (similar) structure of this protein family. As already illustrated in Figure 2, combinations of a-like positions are called patterns or motifs. Position b, however, is not conserved among clusters but it is within each cluster. Thus, position b may be responsible for the functional difference among clusters. In this thesis, we refer to b-like positions as a specificity site. Position c, being highly variable, may neither be important for the overall structure of this protein family nor contribute to the functional differences among subfamilies. The rationale for coining b-like positions as specificity sites is based on the following criteria. First all proteins in the family evolved from a common ancestral sequence. Secondly, they all share a similar structure. Third, the residues on the same positions of the alignment will also overlay in the structures and they probably perform a similar biological function.

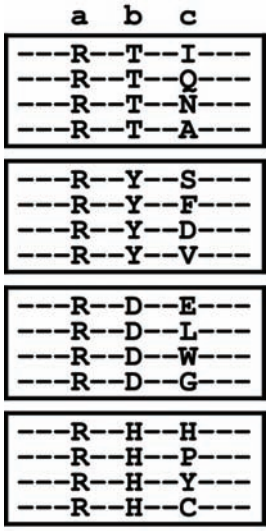


Figure 4 Pseudo sequence alignment of four hypothetic subfamilies of a protein family. Each subfamily has four fictitious sequence fragments. Position a is the conserved position while b is the specificity position which contributes to the functional differences among subfamilies. Position c may not be functionally importance since it is neither conserved between nor within subfamilies.

Information entropy on sequence analysis.

If we consider protein sequences as the language of life, we may use information entropy to measure the information content of a given position in a multiple sequence alignment. In information theory, the **Shannon entropy** or **information entropy** is a measure of the uncertainty associated with a random variable (Shannon, 1948).

The original form of Shannon entropy is

$$H = -\sum_{i=1}^{n} p_i \ln p_i$$

in which n is the number of symbols and p_i is the probability of the i^{th} symbol being chosen (Shannon, 1948). For a given number of symbols, when the probabilities of all symbols are equal, the uncertainty is maximal. Hence the entropy value reaches its maximum value. On the other hand, if the probability of one symbol is 1 while other symbols are forbidden, there is no uncertainty anymore. In this case, the Shannon entropy value will be 0.

In this thesis we use Shannon entropy to quantify the conservation or divergence for a given position in a multiple sequence alignment. F_{ia} is the observed probability of residue a in a given position i of the alignment. Because we have 20 amino acids, we set n to 20.

We let

$$F_{ia} = Number_{ia}/m$$

$$E_i = -\sum_{a=1}^{20} F_{ia} \ln F_{ia}$$

Where $Number_{ia}$ is the number of occurrences of residue a at position i and where m is the number of protein sequences in the alignment. As shown in Figure 5, going from a to e, the positions are more and more divergent because more and more amino acid subtypes join. On position a, there is only one residue, arginine. If we put this information to the above two formulas, the entropy value is 0. However, in position e, there are many different amino acid subtypes so that this position is not conserved at all. If we put this data to the formulas, then we will get a bigger entropy value. One might think of using the *number* of different amino acids to measure conservation on a given position. However, when we compare position b and c, using this approach (account for variability) gives the same value although position b is apparently more conserved than c. The information entropy measure quantifies this difference that relates to the number of different amino acids as well as their frequency distributions.

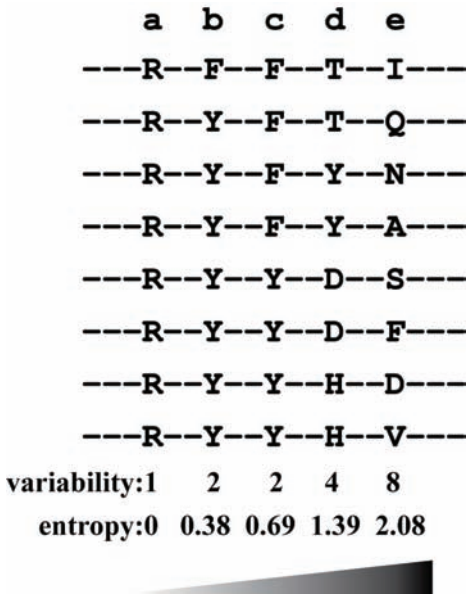


Figure 5 Pseudo sequence alignment of eight hypothetic protein sequences.

Sequential pattern mining

The construction of multiple sequence alignments requires parameterization. It is also computationally intensive, often needs manual adjustment, and can be particularly difficult for a set of deviating sequences. A challenging task is therefore to discover patterns directly from *unaligned* protein sequences. Here we learned from the latest developments in "sequential pattern mining", a new direction in computer science, and adapted a very efficient algorithm, PrefixSpan (Pei et al., 2004), to analyze unaligned protein sequences for common motifs. In computer science, sequential pattern mining allows retrieval of frequently occurring ordered events or subsequences as patterns. An example of a sequential pattern is "Customers who buy a digital camera will probably buy a memory card within one month and a photo printer within another month." Thus the sequential pattern "digital **C**amera" – "**M**emory card" – "photo **P**rinter" may represent certain rules and is commercially valuable for shelf placement and promotions. We may represent the pattern as c*m*p (*: wildcard) since customers may buy other goods in-between.

Protein sequences are sequential arrangements of amino acids by nature. The pattern type presented as c*m*p does not appeal to biologists since both the sequential order of amino acids and the distance (number of divergent residues) between each two conserved residues carries essential biological information. For example, the motif NPxxY is frequently observed in helix 7 of class A G protein-coupled receptors. In Chapter 5 we will further elaborate on the latest developments of sequential pattern mining in computer science and demonstrate how these can be adapted for protein sequence analysis.

Feature selection.

If we consider each protein sequence as an object and the residue at every position of the alignment as a feature, we may borrow techniques of feature selection to identify conserved and specificity positions. Feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is a commonly used technique in machine learning. In this thesis we will use feature selection to identify a subset of relevant features that correlate strongest to the classification. As shown in Figure 4, position b is the correct feature that explains the definition of the four subfamilies.

When we use Shannon entropy to measure conservation within subfamilies and over the entire superfamily, we consider residue positions independently. We developed Multi-RELIEF, a new feature selection algorithm that considers global sequence similarity when searching for specificity residues (see Chapter 4). Mis-classification, a general error that can arise from, e.g., misannotation, will result in classes "polluted" with misplaced sequences. The robustness and accuracy of multi-RELIEF largely prevents this from happening.

Objective and outline of the thesis

The two main objectives of this thesis are i) to develop novel algorithms for the identification of functional positions (conserved or specificity) from either aligned or unaligned protein sequences, and ii) to derive procedures for alignment-independent phylogeny reconstruction.

First, from a given multiple sequence alignment grouped into subfamilies, we use information entropy to measure conservation at the levels of both the entire protein family and its subfamilies (Chapter 2). The "conserved" positions are conserved in both the protein family and subfamilies while "specificity" positions are conserved within subfamilies but divergent among them. The conserved and specificity positions show up on the lower left and upper left corners of the so-called TEA (two-entropies analysis) plot, respectively. However, the definition of protein subfamilies is a challenging problem by itself, particularly because this definition is crucial for the following prediction of specificity positions. In Chapter 3, we automate the method described in Chapter 2 by going through a phylogenetic tree to get a set of subfamily definitions and averaging entire series of TEA plots to yield a consensus TEA-O (two-entropies analysis Objective) plot.

In Chapter 4, we consider the residues in all positions of the multiple sequence alignment as features of the proteins. For a given classification, we developed multi-RELIEF, a machine learning technique for feature weighting, to identify specificity residues. Since specificity positions tend to occur together in a small region, we use such neighboring information from the protein structure to further improve prediction of specificity residues.

In Chapters 2, 3 and 4, we heavily rely on multiple sequence alignments to identify conserved and specificity positions. As mentioned before the construction of such alignments is not self-evident. Following the principles of sequential pattern mining, we propose a new algorithm that directly identifies frequent biologically meaningful patterns from unaligned sequences (Chapter 5).

We were inspired by the procedure of clustering compounds in chemoinformatics and wondered whether we could cluster protein sequences without aligning them. We consider those identified patterns in Chapter 5 as fingerprints of the proteins and use them to calculate a protein distance matrix. This allowed us to construct a phylogenetic tree from such a distance matrix (Chapter 6).

In Chapter 7, conclusions are drawn and further perspectives are discussed.

References

Cuatrecasas, P., Taniuchi, H. and Anfinsen, C.B. (1968) The structural basis of the catalytic function of staphylococcal nuclease, *Brookhaven Symp Biol*, **21**, 172-200.

Gouaux, E. and Mackinnon, R. (2005) Principles of selective ion transport in channels and pumps, *Science*, **310**, 1461-1465.

http://nobelprize.org/.

http://www.nrbsc.org/gfx/genedoc/index.html.

Kozono, D., Ding, X., Iwasaki, I., Meng, X., Kamagata, Y., Agre, P. and Kitagawa, Y. (2003) Functional expression and characterization of an archaeal aquaporin. AqpM from methanothermobacter marburgensis, *J Biol Chem*, **278**, 10649-10656.

Malnic, B., Godfrey, P.A. and Buck, L.B. (2004) The human olfactory receptor gene family, *Proc Natl Acad Sci U S A*, **101**, 2584-2589.

Moore, S. and Stein, W.H. (1973) Chemical structures of pancreatic ribonuclease and deoxyribonuclease, *Science*, **180**, 458-464.

Ngai, J., Chess, A., Dowling, M.M., Necles, N., Macagno, E.R. and Axel, R. (1993) Coding of olfactory information: Topography of odorant receptor expression in the catfish olfactory epithelium, *Cell*, **72**, 667-680.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol*, **302**, 205-217.

Pei, J., Han, J.W., Mortazavi-Asl, B., Wang, J.Y., Pinto, H., Chen, Q.M., Dayal, U. and Hsu, M.C. (2004) Mining sequential patterns by pattern-growth: The PrefixSpan approach, *IEEE Transactions on Knowledge and Data Engineering*, **16**, 1424-1440.

Radzicka, A. and Wolfenden, R. (1995) A proficient enzyme, *Science*, **267**, 90-93.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Mol Biol Evol*, **4**, 406-425.

Shannon, C.E. (1948) A mathematical theory of communication, *The Bell System Technical Journal*, **27**, 54.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res*, **22**, 4673-4680.

Chapter 2

A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors

Motivation: Residues in the transmembrane region of G protein-coupled receptors (GPCRs) are important for ligand binding and activation but the function of individual positions is poorly understood. From one alignment, conserved positions are easily identified and may be important for the folding and the major function. But the ligand binding site in GPCRs is not conserved at all. Can we design an algorithm to detect the ligand binding site directly from a multiple sequence alignment?

Results: Using a sequence alignment of class A GPCRs (grouped in subfamilies), we propose a so-called two-entropies analysis (TEA) to determine the potential role of individual positions in the transmembrane region of class A GPCRs. In our approach, such positions appeared scattered, while largely clustered according to their biological function. Our method appears superior when compared to other bioinformatics approaches such as the evolutionary trace method, entropy-variability plot and correlated mutation analysis, both qualitatively and quantitatively.

Based upon Kai Ye, Eric-Wubbo M. Lameijer, Margot W. Beukers and Adriaan P. IJzerman

PROTEINS: Structure, Function, and Bioinformatics 63:1018–1030 (2006)

Introduction

G protein-coupled receptors (GPCRs) are integral cell membrane proteins that play a crucial role in signal transduction (Schoneberg et al., 2002; Pierce et al., 2002; Gether et al., 2000; Gether et al., 2002). After binding of an endogenous ligand such as a biogenic amine, peptide, nucleotide or even protein, GPCRs undergo a conformational change leading to the activation of heterotrimeric G proteins. GPCRs are very successful drug targets since 30-45% of current drugs interact with this class of proteins (Hopkins and Groom, 2002; Drews, 2000). Consequently, GPCRs represent up to 30% of the portfolio of many pharmaceutical companies (Klabunde and Hessler, 2002).

Despite numerous sequence-function studies on a large number of GPCRs (mainly on class A GPCRs, which represent more than 80% of all GPCRs according to GPCRDB (Horn et al., 2003)), at least two fundamental questions remain. The first is which residues are responsible for the activation mechanism, the second addresses which residues are critical for endogenous ligand binding. Because GPCRs comprise one of the largest superfamilies in the human genome, with almost 1000 proteins (Takeda et al., 2002), various bioinformatics approaches based on multiple sequence alignment have shed light on the above two questions by identifying functional positions, especially at the binding site (Kuipers et al., 1997; Oliveira et al., 2002; Attwood et al., 2002; Oliveira et al., 2003; Attwood et al., 2003; Madabushi et al., 2004; Man et al., 2004). For example, using sequence pattern discovery techniques, Attwood created a database of hierarchical GPCR sequence fingerprints, from superfamily, through family to receptor subtype levels (Kuipers et al., 1997; Attwood et al., 2002; Attwood et., 2003). The fingerprints identified at family-level show a certain correlation to the endogenous ligand binding, whereas the evolutionary trace method has recently been used to reveal global and subfamily-specific conserved residues of class A GPCRs (Madabushi et al., 2004). It was reported that globally conserved residues relate to a canonical conformational switch while some class-specific conserved residues form part of the ligand binding pocket (Madabushi et al., 2004). Correlated mutation analysis and entropy-variability plots were also used to detect networks of functional residues in GPCRs (Oliveira et al., 2002; Oliveira et al., 2003). The combination of these latter two methods allowed the identification of three groups of positions: residues responsible for G protein coupling, residues in the binding site and residues in between these two groups (Oliveira et al., 2002; Oliveira et al., 2003). However, from the same studies it emerges that the resolution, i.e., the capability of unambiguously assigning position to function, is often only modest.

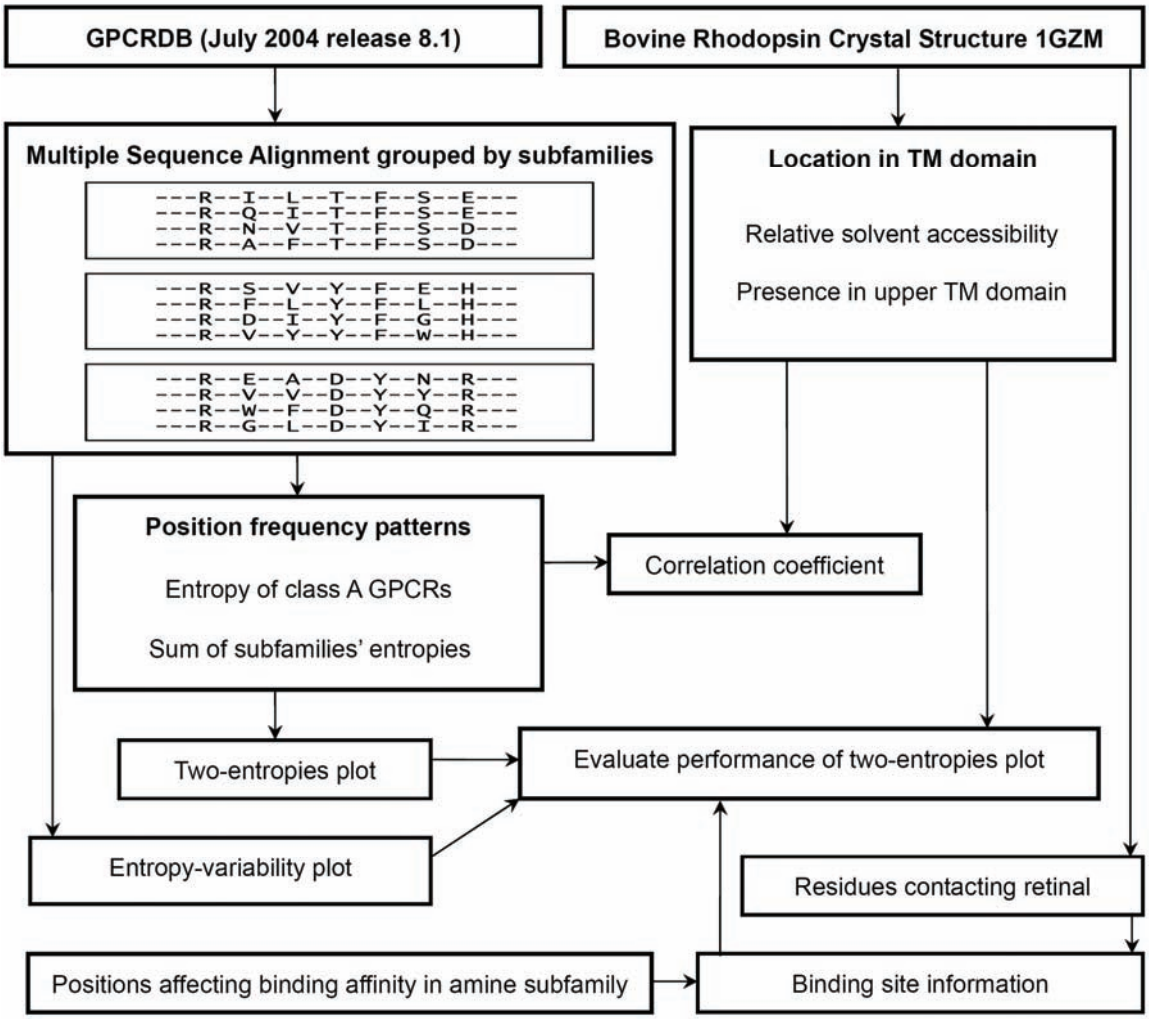


Figure 1. Schematic flowchart of the method used to create and evaluate our two-entropies analysis for class A GPCRs.

In this chapter, we developed a new method, called two-entropies analysis, to identify functional positions of class A GPCRs. The multiple sequence alignment of class A GPCRs was divided into 70 subfamilies based on recognition of identical endogenous ligands (Horn et al., 2003; Horn et al., 1998; Horn et al., 2001; Godfraind et al., 1998). Then, we clustered functional positions based on two observations. The first observation is the variability of each position of class A GPCRs as a whole. The second one is the overall variability of each position within every individual subfamily. We reasoned that positions in a ligand binding site are conserved within subfamilies but divergent among subfamilies. Positions that participate in folding/activation will be largely conserved both within subfamilies and among all class A GPCRs. This principle of predicting positions has been proposed and tried recently using different algorithms on protein families such as protein kinases, bacterial transcription factors

and olfactory receptor (Man et al., 2004; Li et al., 2003; Mirny and Gelfand, 2002; Chiu et al., 2005). However, none of these methods have been applied to a whole set of class A GPCRs.

By comparing our predictions with structural and experimental data as well as other bioinformatics approaches applied to GPCRs (Kuipers et al., 1997; Oliveira et al., 2002; Oliveira et al., 2003; Attwood et al., 2002; Attwood et al., 2003; Madabushi et al., 2004; Man et al., 2004), we were able to provide a global overview of functional positions in the transmembrane region of class A GPCRs. Moreover, our two-entropies analysis proved to be more discriminative than other methods.

Material and Methods

Our approach to cluster positions of class A GPCRs according to their functions and to evaluate the clustering results is depicted in Figure 1, and explained in detail as follows.

Sequences and sequence alignments

All subfamilies of class A GPCRs in the receptor compendium issued by the International Union of Pharmacology (IUPHAR) (Godfraind et al., 1998) were examined in this chapter. The sequence alignment of 1935 class A GPCRs which belong to these subfamilies were extracted from the GPCRDB (July 2004 release 8.1; http://www.gpcr.org/) (Horn et al., 2001). Then according to IUPHAR (Godfraind et al., 1998), the 1935 Class A GPCRs were divided into 70 subfamilies based on recognition of identical endogenous ligands.

Numbering scheme

To facilitate a consistent comparison of aligned residues in different class A GPCRs, we used the indexing method introduced by Ballesteros and Weinstein (1995), in which the most conserved residue in each transmembrane helix is given the index number 50. The other residues are numbered relative to this position.

Definition of boundaries of 7 transmembrane helix regions

The definition of the start and end of the 7 transmembrane helices was adapted from the GPCRDB (Horn et al., 2001). In the numbering scheme of Ballesteros (1995), these boundaries are**:** 1.31 to 1.55 for TM1; 2.42 to 2.66 for TM2; 3.24 to 3.55 for TM3; 4.41 to 4.62 for TM4; 5.38 to 5.58 for TM5; 6.32 to 6.56 for TM6; 7.29 to 7.52 for TM7.

Two-entropies measures

To discriminate amino acid positions that participate in various functions such as binding, and folding/activation, we defined two conservation measures based on the multiple sequence alignment of the entire class A GPCRs and the 70 subfamilies.

Entropy is defined here as a measure of conservation of amino acid residues at a certain position in a defined multiple sequence alignment. The relative frequency F_{ia} of residue type a at alignment position i in a given multiple sequence alignment with m proteins is given by

$$F_{ia} = Number_{ia}/m$$

where $Number_{ia}$ is the number of proteins that have residue type a at alignment position i.

The Shannon entropy at position i in the given alignment is given by equation

$$E_i = -\sum_{a=1}^{20} F_{ia} \ln F_{ia}$$

where a loops over the 20 natural amino acids. We interpret 0ln0 as 0.

Two types of entropies were calculated. The first entropy, the entropy of the entire class A GPCRs, was calculated taking the alignment of all class A GPCRs as input. The second entropy was calculated as the sum of the entropies of all subfamilies of class A GPCRs, while the entropy of each subfamily was calculated taking the individual subfamily sequence alignment as input.

The smaller the entropy value, the more conserved the position is in the given sequence alignment.

Presence in upper or lower domain of the transmembrane region

The positions within the transmembrane region of class A GPCRs were divided into two subsets according to their presence in the upper or lower domain of transmembrane region. The definition of these two domains was adapted from Imai and Fujita (2004). If a position is in the upper domain, a score of 1 was assigned. If a position is in the lower domain, a score of 0 was assigned.

Relative Solvent Accessibility

The solvent accessible surface area of an amino acid residue indicates its level of burial (or solvent exposure) in a protein structure and is often expressed in terms of relative solvent accessibility (RSA). The RSA of a position i of a class A GPCR (RSA_i) was calculated using

the template 1GZM (Li et al., 2004), the crystal structure of bovine rhodopsin. It is defined as the ratio of the solvent exposed surface area of a residue X in position *i* of the bovine rhodopsin crystal structure, denoted as SA_i, and the maximum value of the solvent-exposed surface area for this amino acid corresponding to the surface-exposed area of the central residue observed in the tripeptide GXG in extended conformation, denoted as MSA_i. Thus, RSA_i adopts values between 0% and 100%, with 0% corresponding to a fully buried and 100% to a fully accessible residue, respectively. The computer program MOLMOL 2K.2 (Koradi et al., 1996) was used to compute SA_i, MSA_i, and RSA_i, for positions in the transmembrane region of the crystal structure of bovine rhodopsin 1GZM (Li et al., 2004).

A two-state description distinguishing between residues that are buried (relative solvent accessibility < 15%) and exposed (relative solvent accessibility > 15%) was used (Rost and Sander, 1996).

Correlation matrices of measures

We considered the two forms of entropy described, the relative solvent accessibility, and presence in the upper domain of the transmembrane region as measures for the position's properties. In order to provide a similarity score between these measures, Pearson correlations were performed for each two measures to create correlation matrices, which indicate the distance between every two measures. The Pearson correlation coefficient between any two series of numbers $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$ is defined as

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{\sigma_x}\right)\left(\frac{y_i - \bar{Y}}{\sigma_y}\right)$$

in which $\bar{X}$ and $\bar{Y}$ are the average of these two series of numbers; σ is the standard deviation of these two series of numbers:

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2}$$

The Pearson correlation coefficient is always between -1 and 1, with 1 meaning that two series are perfectly positively correlated, 0 meaning that they are completely uncorrelated, and -1 meaning they are perfectly negatively correlated.

Binding site of bovine rhodopsin based on crystal structure

In order to define the binding site of bovine rhodopsin, the crystal structure 1GZM (Li et al., 2004) was used. Residues within 4 Å distance to the endogenous ligand retinal were considered to be part of the ligand binding site. Calculation was performed with Deepview (Guex and Peitsch, 1997).

Binding sites of aminergic receptors based on experimental data

Information about the binding site of aminergic receptors was derived from Shi and Javitch (2002). Positions were considered to be part of the ligand binding site when they are located in the transmembrane region and implicated in ligand binding in aminergic receptors based on experiments that address affinity labeling, functional complementation of mutations with modifications of ligand, or changes in antagonist affinity.

Entropy-variability plots of class A GPCRs

The graphical representation of our two-entropies analysis is similar to an entropy-variability plot. In order to evaluate the performance of a two-entropies plot in separating positions with different functions, the entropy-variability plots of class A GPCRs were reproduced according to Oliveira et al. (2002) and served as control.

Receiver-operator characteristic (ROC) graph

Receiver-operator characteristic (ROC) graphs provide a visual tool for examining prediction performance (Swets, 1998; Provost and Kohavi, 1998). An ROC graph is a plot with the false positive rate on the x-axis and the true positive rate on the y-axis. It is independent of class distribution or error costs (Provost and Kohavi, 1998).

Two ROC graphs were made to visualize the quantitative comparison of our two-entropies analysis with previous bioinformatics methods in predicting the ligand binding site.

Results

Correlation coefficient between measures

The two types of entropy, the relative solvent accessibility, and the residue's presence in the upper domain of the transmembrane region were used as measures. The correlation coefficients between these measures were calculated and are summarized in Table I.

Some measures were highly correlated, for instance the two types of entropy. This was to be expected because the positions conserved in all class A GPCRs will obviously be conserved in subfamilies too and those that are divergent in proteins within a subfamily will be divergent in the entire class A GPCRs.

Table I Correlation coefficient between measures

	Entropy of class A GPCRs	Presence in the upper domain	Relative solvent accessibility
Sum of subfamilies' entropies	0.678**	0.168*	0.665**
Entropy of class A GPCRs	/	0.401**	0.198**
Presence in upper domain	/	/	-0.042

** Correlation is significant at the 0.01 level (p<0.01);

* Correlation is significant at the 0.05 level (p<0.05)

Separating positions of the upper and lower domain of the TM region

Other correlations are in support of what is known about the sequence-structure relationship in GPCRs as reviewed in the introduction section. For example, the correlation coefficient between entropy of class A GPCRs and the presence of residues in the upper domain of the transmembrane region is 0.401. This correlation coefficient means that the positions in the lower domain are significantly more conserved than those in the upper domain: for the positions in the upper domain (score of presence in the upper domain is 1), the entropy values of class A GPCRs for these positions are generally larger; for the positions in the lower domain (score of presence in the upper domain is 0), the entropy values of class A GPCRs for these positions are largely smaller. The positions in the upper domain involved in ligand binding appear to form a subfamily specific binding site. As for the positions in the lower

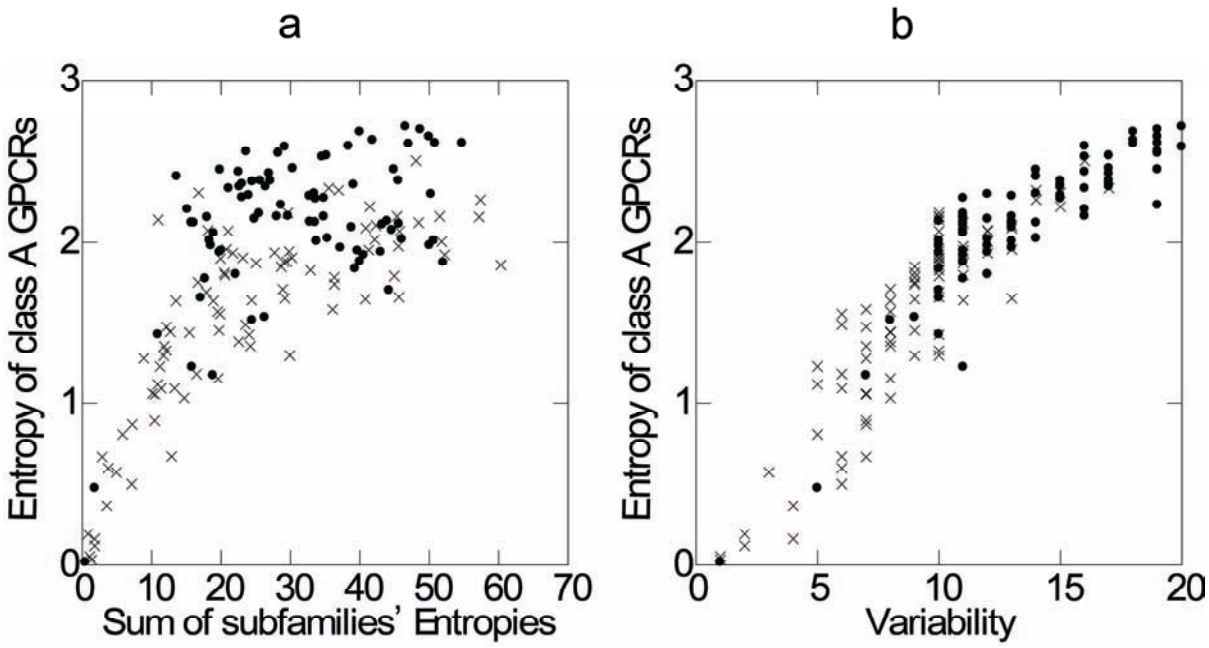


Figure 2. Comparison of the performance of the two-entropies plot and the entropy-variability plot in separating positions in the upper domain and the lower domain of the transmembrane region of class A GPCRs. Dots are positions in upper domain; crosses are positions in lower domain.
a) x-axis is sum of subfamilies' entropies for each position in the transmembrane region; y-axis is entropy of class A GPCRs for each position in the transmembrane region.
b) x-axis is variability for each position in the transmembrane region; y-axis is entropy of class A GPCRs for each position in the transmembrane region.

domain, they are conserved to maintain a similar overall fold and to evoke a similar cascade of activation events.

In Figure 2, we compared the performance of the two-entropies plot versus the entropy-variability plots in separating positions with respect to upper domain (dots) and lower domain (crosses) in the transmembrane region. Both methods illuminate the separation of the two domains.

Separating positions with different relative solvent accessibility

The correlation between relative solvent accessibility and the sum of subfamilies' entropies of all class A GPCRs was more significant than the one between relative solvent accessibility and the entropy of class A GPCRs as a whole (Table I). Apparently, the positions on the surface of the receptors are more divergent than those in the core (blue triangles in Figure 3a). Most positions with large solvent accessibility have higher entropy values for the entire class

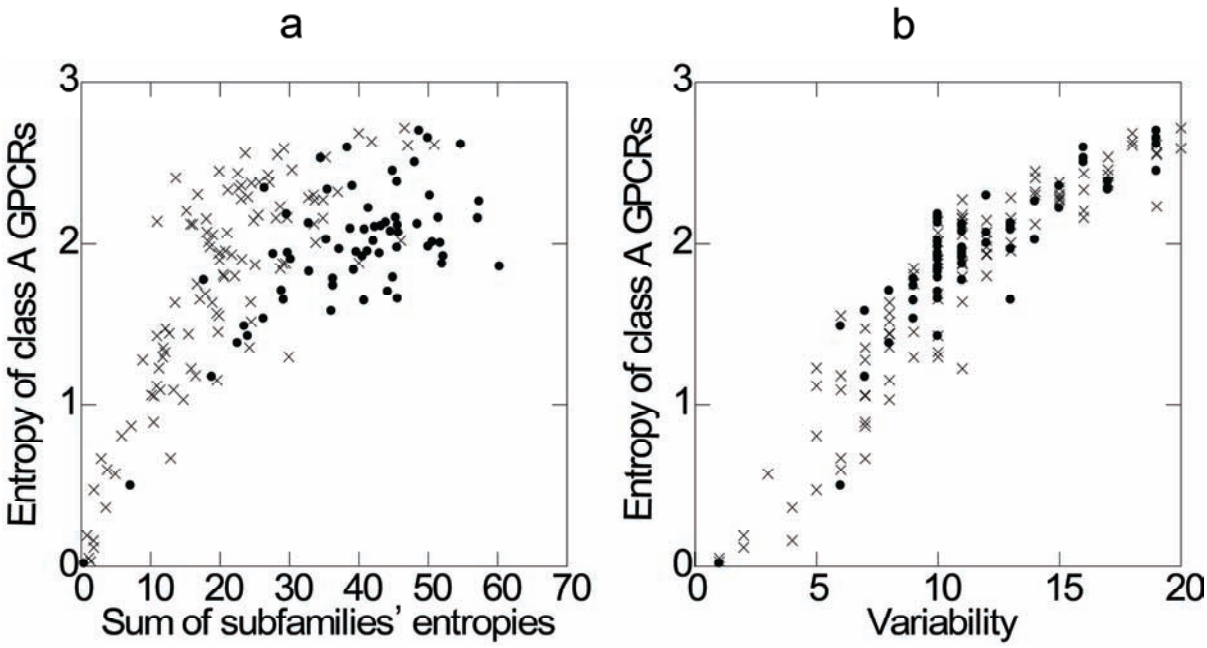


Figure 3. Comparison of the performance of our two-entropies plot and the entropy-variability plot in separating positions with different relative solvent accessibility. Dots are positions with relative solvent accessibility smaller than 15%; crosses are positions with relative solvent accessibility larger than 15%. For labeling of axes, see Figure 2.

A GPCRs. However, dozens of positions in the upper left corner of the two-entropies plot with large entropy values for the entire class A GPCRs (y-axis) and a small sum of the subfamilies' entropies (x-axis) have small solvent accessibility. This suggests that although these positions are in the core of receptors, they are divergent among class A GPCRs but conserved within subfamilies.

The performance of the two-entropies plot (Figure 3a) in separating positions in the core from those on the surface of the transmembrane regions was evaluated and compared with the entropy-variability plot of class A GPCRs (Figure 3b). Although we used the same entropy of class A GPCRs as a measure on the y-axis, the sum of subfamilies' entropies (x-axis) performed better than variability in not only providing a more distinct separation of positions with high variability but also in grouping positions with a similar level of burial in the receptor.

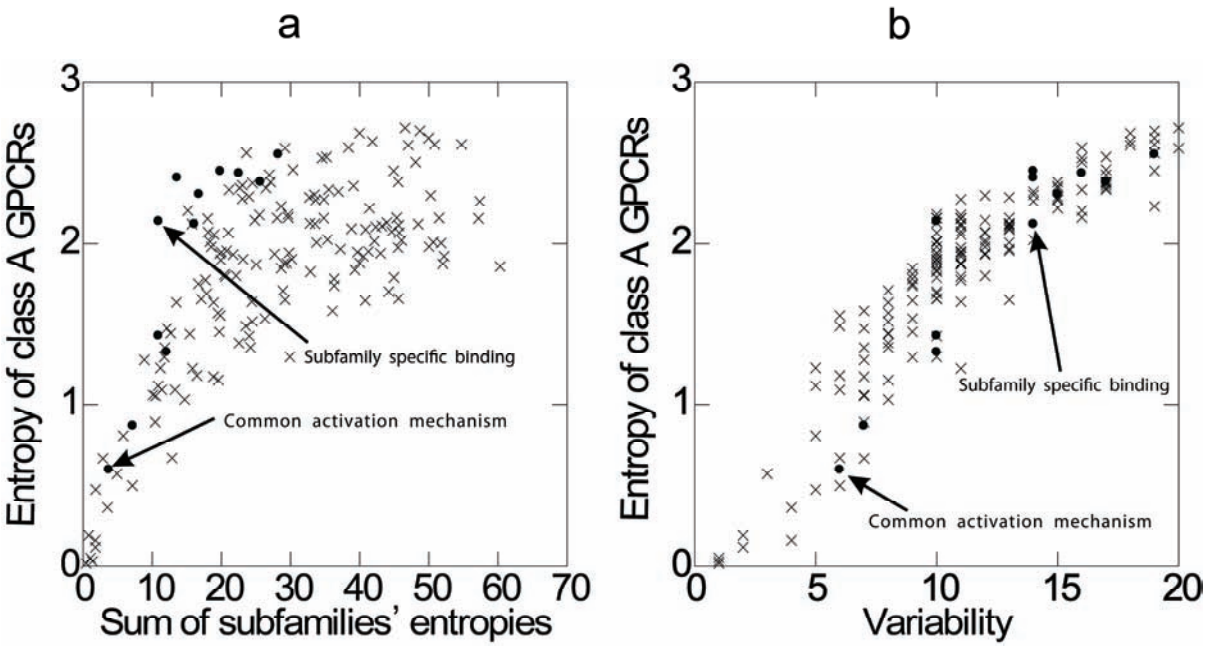


Figure 4. Comparison of the performance of the two-entropies plot and the entropy-variability plot in separating ligand binding site positions from other positions using information derived from the bovine rhodopsin crystal structure 1GZM. Dots are positions within 4 Å distance to retinal, the ligand of bovine rhodopsin; crosses are positions with greater than 4 Å distance to retinal. For labeling of axes, see Figure 2.

Separating positions in the ligand binding site from other positions in the TM region

We collected information about the ligand binding site from both structural and biological data and evaluated the performance of the two-entropies plot in separating positions at the binding site from other positions in the transmembrane region. The binding site of bovine rhodopsin was taken from the crystal structure 1GZM and then mapped onto both the two-entropies plot and the entropy-variability plot as shown in Figure 4. The residues within 4 Å distance to retinal in the crystal structure 1GZM are E113(3.28), A117(3.32), T118(3.33), G121(3.36), E122(3.37), M207(5.42), F212(5.47), F261(6.44), W265(6.48), Y268(6.51), A292(7.39), K296(7.43).

Most positions contacting the ligand of bovine rhodopsin are indeed conserved within subfamilies but show great diversity among different subfamilies (upper left corner of Figure 4a). However, a few positions, such as 5.47, 6.44, 6.48 and 6.51, are conserved with small entropy values with respect to both entropy of class A GPCRs and sum of subfamilies' entropies (lower left corner of Figure 4a). Those conserved positions that contact retinal are exclusively aromatic residues: F212(5.47), F261(6.44), W265(6.48) and Y268(6.51), which

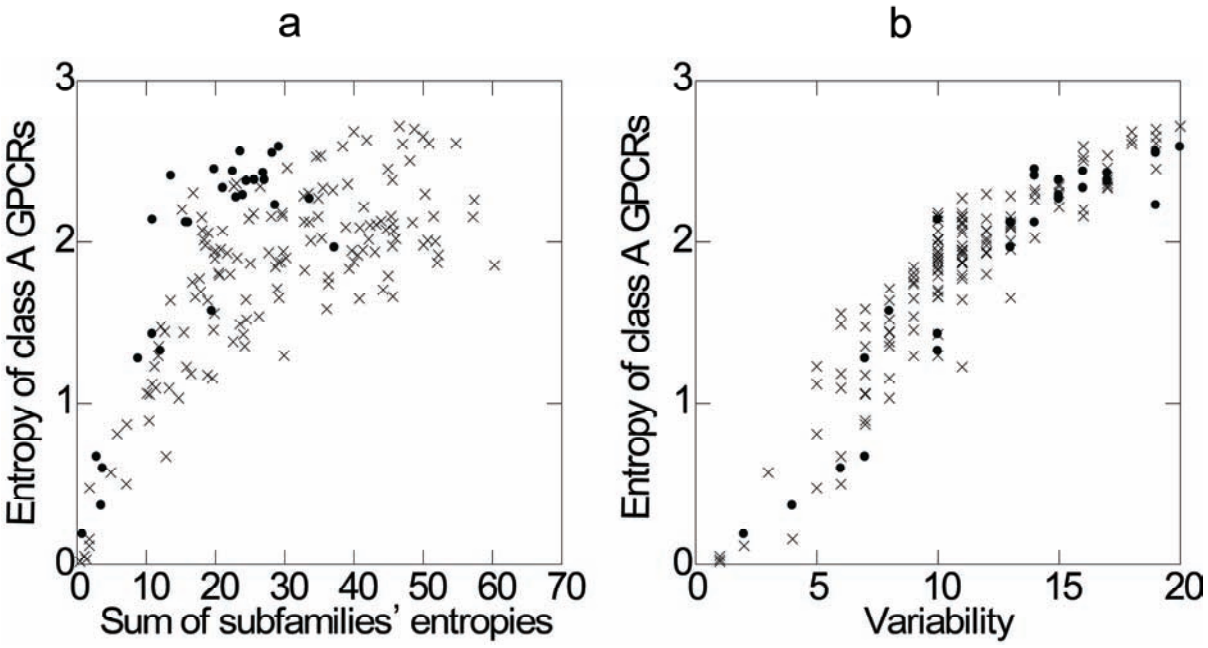


Figure 5. Comparison of the performance of the two-entropies plot and entropy-variability plot in separating binding site positions from other positions. Dots are positions within the transmembrane region implicated in ligand binding in aminergic receptors based on experimental results; Crosses are other positions in the transmembrane region of class A GPCRs. For labeling of axes, see Figure 2.

have been considered as an "aromatic cluster" before by Visier and Ballesteros (2002). According to the authors, once the ligand is recognized by subfamily-specific residues and occupies the binding region, the aromatic cluster will be disturbed and respond through concerted conformational rearrangements of the aromatic side chains to promote receptor activation towards the cytoplasmic side of the receptor. It is safe to conclude that this conserved "aromatic cluster" makes no contribution to the specificity of endogenous ligand binding but that it is responsible for a general activation mechanism (arrow pointing to "common activation mechanism" in Figure 4a).

Thus in the two-entropies plot (Figure 4a), a cluster of positions show up at the upper left corner where the sum of subfamilies's entropies is small and entropy of class A GPCRs is large. These positions probably represent the ligand binding site (arrow pointing to "subfamily specific binding"). However, these positions that may determine subfamily specific binding are mixed with other positions in the entropy-variability plot (Figure 4b).

Positions within the transmembrane region implicated in ligand binding in aminergic receptors based on affinity labeling, functional complementation of mutations with

modifications of ligand, or changes in antagonist affinity (Shi and Javitch, 2002) were mapped onto both the two-entropies plot and the entropy-variability plot. Although very often mutated residues that affect ligand binding are in the ligand binding site, it is also possible that affinity changes are caused by indirect effects such as changing receptor folding or receptor surface expression level (Kristiansen, 2000). In this case, binding site positions derived from the biological data would be distributed more widely in the two-entropies plot (Figure 5a) than compared to the binding site of bovine rhodopsin (Figure 4a). However, most of these positions are still clustered at the upper left corner of the two-entropies plot where we suggest the subfamily-specific binding region to be (Figure 5a).

Similarly as in Figure 4b, the entropy-variability plot (Figure 5b) did not provide a good separation between binding sites and other positions.

Clusters of positions identified by the two-entropies plot

According to Figs. 2-5, positions in the transmembrane region of class A GPCRs in our two-entropies plot tend to cluster according to their functions. After manually mapping positions onto the crystal structure of bovine rhodopsin, we suggest to divide these positions into 6 clusters (Figure 6). The positions in cluster 1 are those that frequently participate in endogenous ligand binding such as position 3.32. These positions in cluster 1 are in the upper domain of the transmembrane region as well as in the core of receptors. They are conserved within subfamilies but divergent among subfamilies. The positions in cluster 2 are involved in folding such as C3.25, which forms a disulfide bridge with a cysteine residue in extracellular loop 2, or in activation such as W6.48 and R3.50. Most positions in cluster 2 are in the lower domain of the transmembrane region and also in the core of receptors. They are conserved among all class A GPCRs. The positions in cluster 3 are at the extracellular end of helices. Among those 21 positions in cluster 3, 4

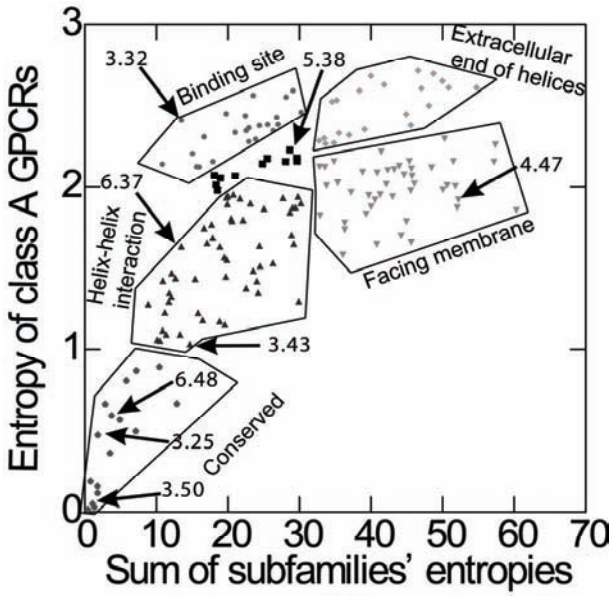


Figure 6. Clustering positions in the two-entropies plot. For labelings of axes, see Figure 2.

positions are at the extracellular end of helix 1; 3 positions are at the extracellular end of helix 2; 7 positions are at the extracellular end of helix 7. The positions in cluster 3 are not conserved within subfamilies and among subfamilies. However, these positions are clustered together in space and they are not far away from the potential binding site. This finding may have potential in drug research, in that synthetic ligands may be modified to contact those positions, to achieve better receptor subtype selectivity. The positions in cluster 4 are slightly less conserved than the 16 positions of cluster 1 with respect to either within subfamilies or among subfamilies. They are located primarily in the lower domain of the transmembrane region and in the core of receptors and are probably involved in helix-helix interaction to conserve the receptor's architecture and to provide a similar activation mechanism for class A GPCRs. Mutation of the positions in cluster 4 can cause receptor constitutive activity, for example positions 3.43 and 6.37 (Lu et al., 1997; Tao et al., 2000; Min and Ascoli, 2000; Latronico et al., 2000; Zeng et al., 1999; Pauwels et al., 2001; Kremer et al., 1999). The positions in cluster 5 are mostly facing the cell membrane. They are divergent both within subfamilies and among all class A GPCRs. However, they are less divergent than positions in cluster 1 with respect to the entire class A GPCRs. Presumably amino acids with various properties will occur in the ligand binding site (cluster 1) of receptors to accommodate variation of ligands in shape, electrostatic and H-bond interactions and aromatic stacking, for example position 3.32 (charged KRHDE 29.63%; aromatic FYW 17.84%; hydrophobic AVLI 27.09%; polar but uncharged STCMNQ 10.04%, G 4.99%, P 1.80%). However for positions facing the membrane, hydrophobic amino acids are more dominant, for instance position 4.47 (charged KRHDE 0.85%; aromatic FYW 2.02%; hydrophobic AVLI 70.49%; polar but uncharged STCMNQ 14.12%, G 10.62%, P 1.80%). The positions in cluster 6 are in the middle of cluster 1, 3, 4 and 5 such that the potential functions of these positions may be a mixture of functions of nearby clusters. For instance, position 5.38, which is close to cluster 1, 3 and 5 is at the end of helix 5 and also at a feasible location to contact the ligand.

Discussion

We divided the functions of positions in the transmembrane region of class A GPCRs into three categories: binding, folding/activation, and "other". Previous studies have shown that strongly conserved positions such as C3.25, R3.50 and W6.48 (numbering scheme according to Ballesteros and Weinstein, 1995) are involved in receptor folding and activation (Gether et al., 2002; Oliveira et al., 2003; Visiers et al., 2002; Kristiansen, 2004; Mirzadegan et al., 2003; Ballesteros et al., 2001; Palczewski et al., 2000). Our approach puts more emphasis on

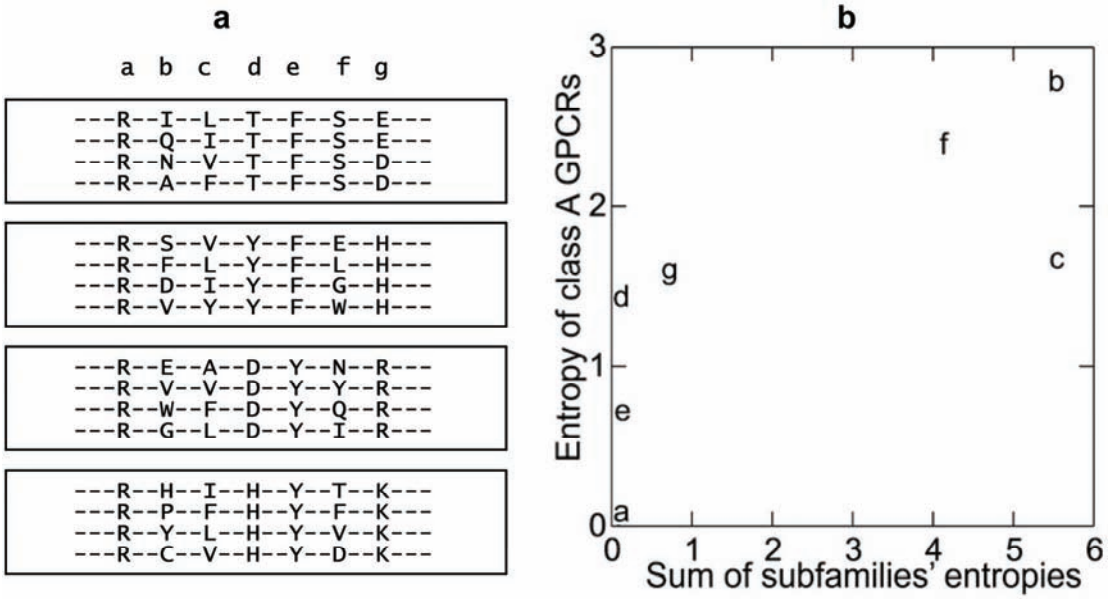


Figure 7. a) Pseudo sequence alignment of four hypothetic subfamilies of class A GPCRs. Each subfamily has four fictitious sequence fragments.
b) Plotting of the positions in a two-entropies plot.

discriminating the binding sites of class A GPCRs from the other two categories. It aims to cluster residues according to their function based on two assumptions. The first is that residues are largely conserved in the binding site of homologous receptors in the same subfamily binding the same endogenous ligand. If so, most GPCRs should share identical residues at binding sites if they belong to the same subfamily and bind an identical endogenous ligand. This is also the unaccounted assumption in both evolutionary trace and correlated mutation analysis during the process of identifying binding sites of GPCRs (Oliveira et al., 2003; Madabushi et al., 2004). The second assumption is that endogenous binding sites are located in a region embedded between transmembrane helices. This has been shown experimentally for a large number of receptors including those for biogenic amines (Liapakis et al., 2000), nucleotides (Jiang et al., 1997), melatonin (Kokkola et al., 2003) and prostacyclin (Stitham et al., 2003).

In the entropy-variability plot, both entropy and variability are measures of conservation of each position of class A GPCRs. The variability at a position is defined as the number of different residue types observed at this position in at least 0.5% of all sequences (Oliveira et al., 2002; Oliveira et al., 2003). Thus the entropy and the variability are strongly correlated and positions in entropy-variability plots are crowded along the diagonal. For this reason, in our two-entropies plot we only adopted one measure, entropy of all class A GPCRs, in order

to separate overall conserved positions from divergent positions. The second measure, sum of subfamilies' entropies, was introduced to scatter positions with high entropy values of all class A GPCRs. In this way, our two-entropies plot achieves a better balance of robustness and sensitivity than the entropy-variability plot or the evolutionary trace method, which will be discussed later.

In order to compare the performance of our two-entropies plot with other sequence alignment-based methods in differentiating positions with different functions, we give a sequence alignment for four hypothetical subfamilies of class A GPCRs. Each subfamily is suggested to bind a different endogenous ligand. The alignment was established within subfamilies and also between subfamilies (Figure 7a). Note that this hypothetical set of GPCR sequences was only used to illustrate the principle of our approach. As described in the results section, our further analysis was based on the total set of 1935 GPCR sequences from 70 subfamilies.

The two entropies of each position in Figure 7a were calculated according to the algorithm described in the Methods section and are shown in Table II. Note that due to a relatively small number of subfamilies (4 subfamilies, Figure 7a) and either perfect conservation or divergence in the sequence alignment (Figure 7a), the overall configuration of positions in Figure 7b is more outspoken than in Figs. 2-6. For example, in a more realistic situation, position *d* indeed has a small sum of subfamilies' entropies but not zero. Its entropy value of all subfamilies will be larger because many more subfamilies (70) are present in the alignment than the 4 hypothetic ones.

Table II. Entropy of sequence alignment and sum of subfamilies' entropies of sequence alignment in Figure 7a

	a	*b*	*c*	*d*	*e*	*f*	*g*
Entropy of all subfamilies	0	2.77	1.66	1.39	0.69	2.39	1.56
Sum of subfamilies' entropies	0	5.55	5.55	0	0	4.16	0.69

Our two-entropies plot (Figure 7b) illustrates the differences between positions *a*, *b*, *c*, *d*, *e*. For position *a*, amino acids are conserved within and between subfamilies. Both entropies are small and this position *a* will appear in the lower left corner of the two-entropies plot.

Functions of position *a* could be folding such as C3.25, which forms a disulfide bridge with a conserved cysteine residue in extracellular loop 2, or activation such as W6.48 and R3.50.

As for position *b*, amino acids are neither conserved between subfamilies nor conserved in the individual subfamilies. In addition, hydrophilic, hydrophobic and aromatic residues show up in position *b*. In this case, both entropy measures will have large values. The function of position *b* may be other than ligand binding and folding/activation.

Position *c* is quite similar to position *b*, amino acids are neither conserved between subfamilies nor conserved in each subfamily. However, only hydrophobic residues show up in position *c*. So, both entropy measures have large values but smaller than those of position *b*. Residues in position *c* probably face the membrane.

Position *d* is very important. Although all 20 residues amino acids may show up at this position, they are mostly conserved within each subfamily but divergent among subfamilies. For position *d*, the sum of subfamilies' entropies will be small and the entropy of class A GPCRs will be large. It is probable that position *d* frequently participates in endogenous ligand binding.

Position *e* is also very important. Residues are conserved within each subfamily and also shared by several subfamilies. As a result, the sum of subfamilies' entropies will be small but the entropy of class A GPCRs will be larger than position *a* and smaller than position *b*, *c* and *d*. Position *e* may participate in helix-helix interactions to conserve the 3D aspects of GPCRs and to provide a common activation mechanism for class A GPCRs.

Positions *f* and *g* will be discussed later to compare the evolutionary trace method with our two-entropies analysis.

Two-entropies plot versus entropy-variability plot

Although entropy-variability plots have been used very successfully in the past (Oliveira et al., 2002; Oliveira et al., 2003), the performance of our two-entropies plot in separating positions according to their functions appears improved (Figure 2-5). Most importantly, as shown in Figure 4 and Figure 5, the entropy-variability plot does not differentiate very well between positions *b* and *d*. The explanation is as follows. When more than 20 subfamilies are present in one superfamily, 20 residues types are likely to be present in each position of the binding region to account for the diversity of endogenous ligands. This is the case for the large family of class A GPCRs, and hence the entropy value of position *b* will be as large as the one of position *d* and the variability of both positions *b* and *d* will be close to 20.

Two-entropies plot versus the evolutionary trace method

The evolutionary trace method has been shown successful in predicting binding sites of soluble and membrane proteins (Zhu et al., 2004; Shackelford et al., 2004; Blaise et al., 2004; Innis et al., 2000; Xie et al., 1999; Pritchard and Dufton et al., 1999; Lichtarge et al., 1996). Recently, this method has been used to analyze class A GPCR sequences to identify globally conserved residues and opsin subfamily-specific residues (Madabushi et al., 2004). In that study, only four subfamilies of class A GPCRs, visual opsin, bioamine, olfactory, and chemokine, were included to trace 39 "globally" conserved residues. Only the opsin subfamily was subjected to differential trace analysis and finally 17 opsin "specific" conserved residues were identified (Madabushi et al., 2004).

However, the identified 39 "globally" conserved residues based on only four subfamilies are not conserved in all subfamilies of class A GPCRs. For example, position 3.33 was identified as one of the 39 "globally" conserved residues. But great variation in position 3.33 is observed among all class A GPCRs (Figure 8). Because it is hard to include dozens of subfamilies in the evolutionary trace method and to compare subfamily-specific conserved residues between every pair of subfamilies, we believe the evolutionary trace method does not make full use of the rich sequence information of a superfamily as large as class A GPCRs.

In addition, the evolutionary trace method does not easily recognize positions with small variation as globally conserved residues. For example, the method failed to identify position 2.50 as a globally conserved position (Madabushi et al., 2004), which has D in 92% of class A GPCRs. Suppose among 1000 proteins of class A GPCRs, only two amino acid types (e.g., D and K) are present in a certain position. Obviously, there is a great difference between a situation with 1D/999K versus 500D/500K. Unfortunately, the evolutionary trace method ignores such a difference and considers both of the above two situations as a non-conserved position. However, both the entropy-variability plot and our two-entropies plot detect such a conservation because they are designed to measure conservation on the basis of not only the number of amino acid types at a given position but also the frequency of each amino acid type at that position.

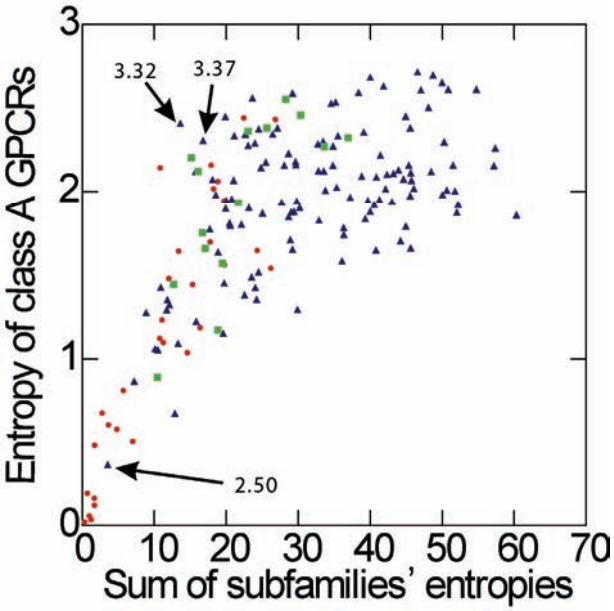


Figure 8. Plotting "globally" conserved positions and opsin subfamily "specific" conserved positions (Madabushi et al., 2004) identified by the evolutionary trace method onto our two-entropies plot. x-axis is sum of subfamilies' entropies; y-axis is entropy of class A GPCRs. Red dots are "globally" conserved positions; green squares are opsin subfamily "specific" conserved positions; blue triangles are other positions.

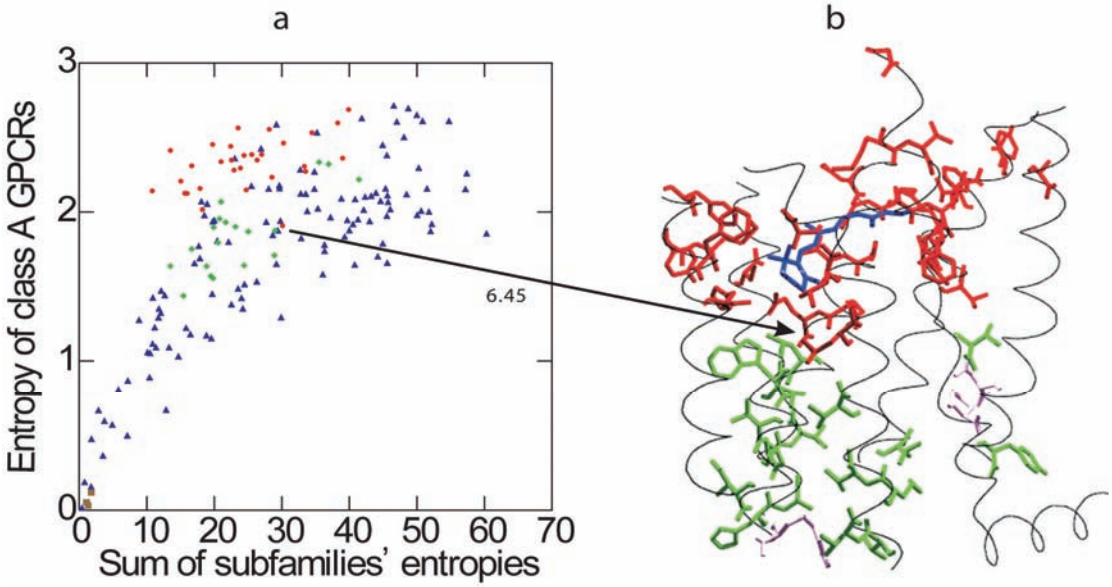


Figure 9. Plotting three networks of positions identified by CMA (Oliveira et al., 2002) in the two-entropies plot (a) and mapping these three networks onto the crystal structure of bovine rhodopsin (b). Pink squares are conserved positions (network 1). Red dots are involved in ligand binding (network 2). Green rhombuses are involved in G protein coupling and activation (network 3).

In principle, the evolutionary trace method differentiates between positions *a*, *b*, *d* and *e* (Figure 7). However, it may make mistakes at position *f* which is conserved in just one subfamily, and hence considers position *f* as in the ligand binding site. The position *f* may not be functionally important because it is possible that such "conservation" is caused by a small

population of proteins or a short evolution history since the subfamily member began to evolve. In our two-entropies plot, position *f* is not misjudged because our approach does not take just one subfamily into account but the overall observation in all subfamilies.

The evolutionary trace method may also lead to erroneous results in subfamilies in which ligand binding sites are not completely conserved such as position *g* (Figure 7a). For instance, in adenosine receptors, position 7.42 was reported to be involved in agonist binding (Townsend-Nicholson and Schofield, 1994; Tucker et al., 1994; Kim et al., 1995; Jiang et al., 1996; Dalpiaz et al., 1998). However, position 7.42 is a serine in the human adenosine A_{2A} receptor but a threonine in the human adenosine A_1 receptor. Because of the high sensitivity to class-specific conservation, any slight variation at the binding site will impede the evolutionary trace method in identifying the binding site. In contrast, our two-entropies plot will still consider position *g* as belonging to the binding site, since the joint conservation within subfamilies and large divergence in all class A GPCRs strongly indicates that this position is involved in the ligand binding. For this reason, the evolutionary trace method probably failed to predict two positions, 3.32 and 3.37, as part of the binding site of bovine rhodopsin. However, these two positions are located at the upper left corner of the two-entropies plot and they are indeed within 4 Å distance to retinal, the endogenous ligand of bovine rhodopsin.

Two-entropies plot versus sequence pattern discovery

Various sequence pattern discovery approaches have been applied to GPCRs. For example, using sequence pattern discovery techniques, Attwood created a database of hierarchical GPCR sequence fingerprints, from superfamily, through family to receptor subtype levels (Kuipers et al., 1997; Attwood et al., 2002; Attwood et al., 2003). The fingerprints identified at family-level show a certain correlation to the endogenous ligand binding.

Compared to the sequence pattern discovery approaches, our approach predicts the functional sites of GPCRs in a more precise way for two reasons. First of all, our method exploits the conservation among all subfamilies rather than per subfamily. Second, out method can handle very large numbers of sequences at the same time. In contrast, sequence pattern discovery algorithms investigate only one phylogenetic level of subfamily each time without a cross-check between different subfamilies. This defect does harm as we can see in the evolutionary trace method discussed above: some positions in the identified sequence patterns may be conserved in a certain subfamily, but they may not be functionally important because such conservation is caused by the small population of the subfamily or the short evolution

history since the subfamily member began to evolve. Thus, such approaches never take full advantage of the huge superfamily with dozens of subfamilies such as GPCRs.

Two-entropies plot versus correlated mutation analysis

The method of correlated mutation analysis (CMA) was proposed to detect intramolecular or intermolecular contacts or links between residues. It successfully predicted the approximate binding region of class A GPCRs (Oliveira et al., 2002; Oliveira et al., 2003). The principle of CMA in defining the binding region of class A GPCRs is that when one endogenous ligand changes to another, residues in the binding site will "mutate" at the same time to accommodate the change of ligand in shape, and in hydrophobic and electrostatic properties. Correlated mutation analysis easily discriminates position *a* and position *e* in the pseudo alignment shown in Figure 7a. As shown in Figure 9a, CMA is consistent with our two-entropies plot: most binding sites of class A GPCRs predicted by CMA are clustered at the upper left corner of our two-entropies plot.

However, let us differentiate the residues close to the binding site into two layers. The first layer of residues surrounds the ligand so that they are indeed at the binding site. The residues contacting the first layer of residues but not the ligand constitute the second layer of residues. When the first-layer residues "mutate" to accommodate a different ligand, the second-layer residues will also "mutate" at the same time to maintain compact contact and correct interaction with the first layer. The correlation between the first and the second layer may be so strong that it is hard to discriminate the first layer from the second. For this reason, prediction of binding sites of class A GPCRs by CMA also includes positions facing the membrane such as position 6.45 (Figure 9). In addition, CMA failed to predict positions 5.39, 2.64 and 4.60 as involved in ligand binding (Figure 9), mutation of which affected agonist binding affinity in aminergic receptors (Shi et al., 2002). These three positions, however, are conserved within subfamilies and divergent between subfamilies so that they are indeed part of the binding site of class A GPCRs according to the two-entropies plot. Finally, CMA will only detect a functional network where positions are either conserved or strongly correlated with other positions. For example, in Figure 9a, many positions in the lower left corner of the two-entropies plot are "invisible" to CMA because the frequencies of the 20 amino acids in these positions show no correlation with other positions, although these positions are functionally important.

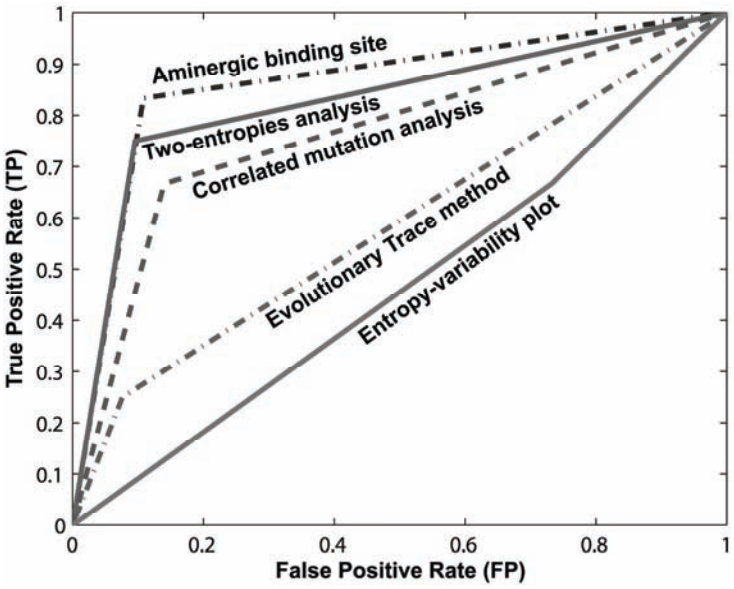


Figure 10. ROC graph for the prediction of the ligand binding site of bovine rhodopsin. The true positive rate and the false positive rate were calculated for each bioinformatics approach for the prediction of the ligand binding site of bovine rhodopsin, using its crystal structure as "golden standard".

Two-entropies plot versus "Mutual Information"

Mirny and coworkers used a "Mutual Information" approach to measure both conservation within subfamily and diversity between subfamilies. The authors used various statistical models to evaluate significance of mutual information and to identify so-called "specificity-determining" positions (Li et al., 2003; Mirny and Gelfand, 2002).

In our approach, we use two measures to evaluate the conservation of positions among proteins within the same subfamily and their diversity between different subfamilies. The combined two measures (*entropies*) result in a high resolution in identifying binding sites and other functional sites.

Although our method apparently provides a global overview of all functional positions, it is not known yet how well the two methods of two-entropies analysis and mutual information compare, since they have not been applied to the same dataset. Thus we are currently applying our method to the superfamily of protein kinases, which have already been analyzed by mutual information (Li et al., 2003).

Quantitative comparison with previous bioinformatics methods

Two receiver-operator characteristic (ROC) plots (Provost and Kohavi, 1998) were made to visualize the quantitative comparison of our two-entropies analysis with previous bioinformatics methods.

As mentioned above, the ligand binding site (12 positions) of bovine rhodopsin was taken from the crystal structure 1GZM. Our two-entropies analysis and other bioinformatics

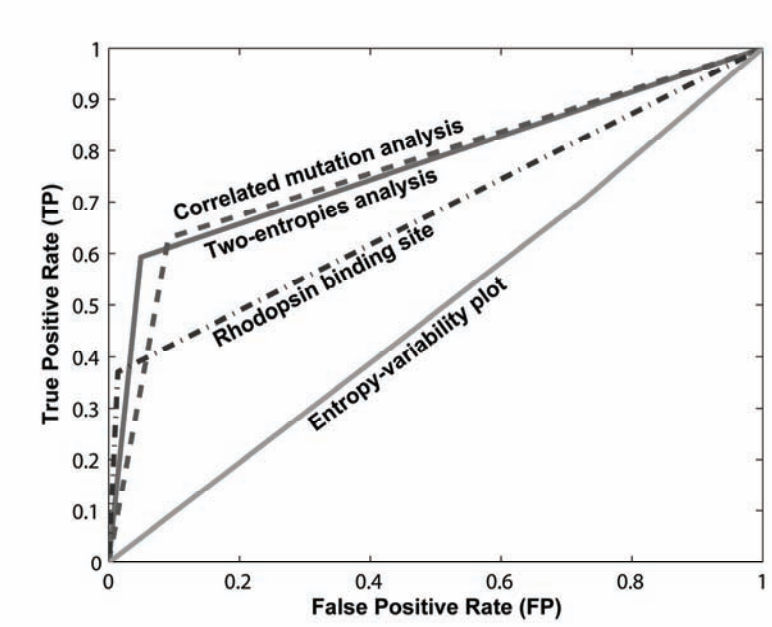


Figure 11. ROC graph for the prediction of the ligand binding site of aminergic receptors. The true positive rate and the false positive rate were calculated for each bioinformatics approach for the prediction of the ligand binding site of aminergic receptors, using the experimental data as "golden standard".

methods were compared using an ROC plot (Figure 10). The binding site information of aminergic receptors based on experimental data was also included in this comparison. Apparently, our two-entropies analysis performs better than other bioinformatics methods. If the binding site information of aminergic receptors based on experimental data was used to predict the binding site of rhodopsin, it performed slightly better (correctly predicted 10 out of 12 amino acids) than our two-entropies analysis (predicted 9 of 12 residues). However, the three positions (F5.47, F6.44, W6.48) that our method missed are conserved among class A GPCRs and belong to the highly conserved aromatic cluster, which does not contribute to subfamily-specific ligand binding, but instead part of the activation cascade. In other words, our method successfully predicted all subfamily-specific residues of bovine rhodopsin that determine endogenous ligand binding and was able to discriminate between ligand binding residues and activation residues.

The positions involved in rhodopsin ligand binding are only a subset of the general ligand binding positions in class A GPCRs as predicted by our two-entropies analysis. The ligand binding residues that we have identified form a larger set of residues to accommodate the various sizes and shapes of the endogenous ligands of class A GPCRs. Thus, each endogenous ligand binds to a subset of these residues and the binding site determined from the bovine rhodopsin crystal structure represents the residues that make up the retinal binding site. As a result, the binding site determined from the bovine rhodopsin crystal structure cannot well represent other class A GPCRs as shown in the next ROC plot (Figure 11).

As mentioned above, the information about the binding site of aminergic receptors was derived from Shi and Javitch (2002). When the binding site information of bovine rhodopsin based on its crystal structure was used to predict the binding site of aminergic receptors, it performed much worse than our two-entropies analysis and correlated mutation analysis. Apparently, rhodopsin and aminergic receptors use different subsets of the general binding site as their ligand binding site.

Conclusion

Based on the sequence alignment of class A GPCRs grouped into subfamilies, a two-entropies analysis is proposed to determine the potential functions of positions in the transmembrane region of GPCRs. In our two-entropies plot approach, positions of class A GPCRs in the transmembrane region were scattered and clustered according to their biological functions. The two-entropies analysis may also be applicable to other protein superfamilies.

References

Attwood TK. (2002) The PRINTS database: A resource for identification of protein families, Brief Bioinform, **3**: 252-263.

Attwood TK, Blythe MJ, Flower DR, Gaulton A, Mabey JE, Maudling N, McGregor L, Mitchell AL, Moulton G, Paine K, Scordis P. (2002) PRINTS and PRINTS-S shed light on protein ancestry, Nucleic Acids Res, **30**: 239-241.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C. (2003) PRINTS and its automatic supplement, preprints, Nucleic Acids Res, **31**: 400-402.

Ballesteros JA, Shi L, Javitch JA. (2001) Structural mimicry in G protein-coupled receptors: Implications of the high-resolution structure of rhodopsin for structure-function analysis of rhodopsin-like receptors, Mol Pharmacol, **60**: 1-19.

Ballesteros JA, Weinstein, H. (1995) Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors, Methods in Neurosciences, **25**: 366-428.

Blaise MC, Sowdhamini R, Rao MR, Pradhan N. (2004) Evolutionary trace analysis of ionotropic glutamate receptor sequences and modeling the interactions of agonists with different NMDA receptor subunits, J Mol Model, **10**: 305-316

Chiu HC, Chang CA, Hu YJ. (2005) Prediction of orthologous relationship by functionally important sites, Comput Methods Programs Biomed, **78**: 209-222.

Dalpiaz A, Townsend-Nicholson A, Beukers MW, Schofield PR, IJzerman AP. (1998) Thermodynamics of full agonist, partial agonist, and antagonist binding to wild-type and mutant adenosine A_1 receptors, Biochem Pharmacol, **56**: 1437-1445.

Drews J. (2000) Drug discovery: A historical perspective, Science **287**: 1960-1964.

Pierce KL, Premont RT, Lefkowitz RJ. (2002) Seven-transmembrane receptors, Nat Rev Mol Cell Biol **3**: 639-650.

Gether U, Asmar F, Meinild AK, Rasmussen SG. (2002) Structural basis for activation of G-protein-coupled receptors, Pharmacol Toxicol, **91**: 304-312.

Gether U. (2000) Uncovering molecular mechanisms involved in activation of G protein-coupled receptors, Endocr Rev, **21**: 90-113.

Godfraind T, Vanhoutte P, Ruffolo R, Humphrey P. (1998) The IUPHAR Compendium of Receptor Characterization and Classification, London: IUPHAR Media; 267.

Guex N, Peitsch MC. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling, Electrophoresis, **18**: 2714-2723.

Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G. (2003) GPCRDB information system for G protein-coupled receptors, Nucleic Acids Res, **31**: 294-297.

Horn F, Vriend G, Cohen FE. (2001) Collecting and harvesting biological data: The GPCRDB and NucleaRDB information systems, Nucleic Acids Res, **29**: 346-349.

Horn F, Weare J, Beukers MW, Horsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G, (1998) GPCRDB: An information system for G protein-coupled receptors. Nucleic Acids Res, **26**: 275-279.

Hopkins AL, Groom CR, (2002) The druggable genome, Nat Rev Drug Discov, **1**: 727-730.

Imai T, Fujita N. (2004) Statistical sequence analyses of G-protein-coupled receptors: Structural and functional characteristics viewed with periodicities of entropy, hydrophobicity, and volume. Proteins, **56**: 650-660.

Innis CA, Shi J, Blundell TL. (2000) Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis, Protein Eng, **13**: 839-847.

Jiang Q, Guo D, Lee BX, Van Rhee AM, Kim YC, Nicholas RA, Schachter JB, Harden TK, Jacobson KA. (1997) A mutational analysis of residues essential for ligand recognition at the human P2Y1 receptor, Mol Pharmacol, **52**: 499-507.

Jiang Q, Van Rhee AM, Kim J, Yehle S, Wess J, Jacobson KA. (1996) Hydrophilic side chains in the third and seventh transmembrane helical domains of human A_{2A} adenosine receptors are required for ligand recognition, Mol Pharmacol, **50**: 512-521.

Klabunde T, Hessler G. (2002) Drug design strategies for targeting G-protein-coupled receptors. Chembiochem, **3**: 928-944.

Koradi R, Billeter M, Wuthrich K. (1996) MOLMOL: A program for display and analysis of macromolecular structures, J Mol Graph, **14**:51-55, 29-32.

Kremer H, Martens JW, van Reen M, Verhoef-Post M, Wit JM, Otten BJ, Drop SL, Delemarre-van de Waal HA, Pombo-Arias M, De Luca F, Potau N, Buckler JM, Jansen M, Parks JS, Latif HA, Moll GW, Epping W, Saggese G, Mariman EC, Themmen AP, Brunner HG. (1999) A limited repertoire of mutations of the luteinizing hormone (LH) receptor gene in familial and sporadic patients with male LH-independent precocious puberty, J Clin Endocrinol Metab, **84**: 1136-1140.

Kristiansen K. (2004) Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function, Pharmacol Ther, **103**: 21-80.

Kristiansen K, Kroeze WK, Willins DL, Gelber EI, Savage JE, Glennon RA, Roth BL. (2000) A highly conserved aspartic acid (Asp-155) anchors the terminal amine moiety of tryptamines and is involved in membrane targeting of the 5-HT(2A) serotonin receptor but does not participate in activation via a "salt-bridge disruption" mechanism, J Pharmacol Exp Ther, **293**: 735-746.

Kuipers W, Oliveira L, Vriend G, IJzerman AP. (1997) Identification of class-determining residues in G protein-coupled receptors by sequence analysis, Receptors Channels, **5**: 159-174.

Kim J, Wess J, van Rhee AM, Schoneberg T, Jacobson KA. (1995) Site-directed mutagenesis identifies residues involved in ligand recognition in the human A_{2A} adenosine receptor, J Biol Chem, **270**: 13987-13997.

Kokkola T, Foord SM, Watson MA, Vakkuri O, Laitinen JT. (2003) Important amino acids for the function of the human MT1 melatonin receptor, Biochem Pharmacol, **65**: 1463-1471.

Lichtarge O, Bourne HR, Cohen FE. (1996) An evolutionary trace method defines binding surfaces common to protein families, J Mol Biol, **257**: 342-358.

Latronico AC, Shinozaki H, Guerra G, Jr., Pereira MA, Lemos Marini SH, Baptista MT, Arnhold IJ, Fanelli F, Mendonca BB, Segaloff DL. (2000) Gonadotropin-independent precocious puberty due to luteinizing hormone receptor mutations in Brazilian boys: A novel constitutively activating mutation in the first transmembrane helix, J Clin Endocrinol Metab, **85**: 4799-4805.

Li J, Edwards PC, Burghammer M, Villa C, Schertler GF. (2004) Structure of bovine rhodopsin in a trigonal crystal form, J Mol Biol, **343**: 1409-1438.

Li L, Shakhnovich EI, Mirny LA. (2003) Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases, Proc Natl Acad Sci U S A, **100**: 4463-4468.

Lu ZL, Curtis CA, Jones PG, Pavia J, Hulme EC. (1997) The role of the aspartate-arginine-tyrosine triad in the M1 muscarinic receptor: Mutations of aspartate 122 and tyrosine 124 decrease receptor expression but do not abolish signalling, Mol Pharmacol, **51**: 234-241.

Lu ZL, Hulme EC. (1999) The functional topography of transmembrane domain 3 of the M1 muscarinic acetylcholine receptor, revealed by scanning mutagenesis, J Biol Chem, **274**: 7309-7315.

Liapakis G, Ballesteros JA, Papachristou S, Chan WC, Chen X, Javitch JA. (2000) The forgotten serine. A critical role for Ser-203 5.42 in ligand binding to and activation of the beta 2-adrenergic receptor, J Biol Chem, **275**: 37779-37788.

Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O. (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions, J Biol Chem, **279**: 8126-8132.

Man O, Gilad Y, Lancet D. (2004) Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons, Protein Sci, **13**: 240-254.

Min L, Ascoli M. (2000) Effect of activating and inactivating mutations on the phosphorylation and trafficking of the human lutropin/choriogonadotropin receptor, Mol Endocrinol, **14**: 1797-1810.

Mirny LA, Gelfand MS. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors, J Mol Biol, **321**: 7-20.

Mirzadegan T, Benko G, Filipek S, Palczewski K. (2003) Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin, Biochemistry, **42**: 2759-2767.

Oliveira L, Paiva AC, Vriend G. (2002) Correlated mutation analyses on very large sequence families, Chembiochem, **3**: 1010-1017.

Oliveira L, Paiva PB, Paiva AC, Vriend G. (2003) Identification of functionally conserved residues with the use of entropy-variability plots, Proteins, **52**: 544-552.

Oliveira L, Paiva PB, Paiva AC, Vriend G. (2003) Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein, Proteins, **52**: 553-560.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor, Science. **289**: 739-745.

Pritchard L, Dufton MJ. (1999) Evolutionary trace analysis of the Kunitz/BPTI family of proteins: Functional divergence may have been based on conformational adjustment, J Mol Biol, **285**: 1589-1607.

Pauwels PJ, Wurch T, Tardif S, Finana F, Colpaert FC. (2001) Analysis of ligand activation of alpha 2-adrenoceptor subtypes under conditions of equal G alpha protein stoichiometry, Naunyn Schmiedebergs Arch Pharmacol, **363**: 526-536.

Provost F, Kohavi R. (1998) Guest editors' introduction: On applied research in machine learning. Machine Learning, **30**: 127-132.

Rost B, Sander C. (1996) Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct, **25**: 113-136.

Takeda S, Kadowaki S, Haga T, Takaesu H, Mitaku S. (2002) Identification of G protein-coupled receptor genes from the human genome sequence, FEBS Lett, **520**: 97-101.

Tao YX, Abell AN, Liu X, Nakamura K, Segaloff DL. (2000) Constitutive activation of G protein-coupled receptors as a result of selective substitution of a conserved leucine residue in transmembrane helix III, Mol Endocrinol, **14**: 1272-1282.

Townsend-Nicholson A, Schofield PR. (1994) A threonine residue in the seventh transmembrane domain of the human A_1 adenosine receptor mediates specific agonist binding, J Biol Chem, **269**: 2373-2376.

Tucker AL, Robeva AS, Taylor HE, Holeton D, Bockner M, Lynch KR, Linden J. A_1 adenosine receptors. (1994) Two amino acids are responsible for species differences in ligand recognition, J Biol Chem, **269**: 27900-27906.

Shi L, Javitch JA. (2002) The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop, Annu Rev Pharmacol Toxicol, **42**: 437-467.

Swets JA. (1988) Measuring the accuracy of diagnostic systems, Science, **240**: 1285-1293.

Schoneberg T, Schulz A, Gudermann T. (2002) The structural basis of G-protein-coupled receptor function and dysfunction in human diseases, Rev Physiol Biochem Pharmacol, **144**: 143-227.

Shackelford GS, Regni CA, Beamer LJ. (2004) Evolutionary trace analysis of the alpha-D-phosphohexomutase superfamily, Protein Sci, **13**: 2130-2138.

Stitham J, Stojanovic A, Merenick BL, O'Hara KA, Hwa J. (2003) The unique ligand-binding pocket for the human prostacyclin receptor. Site-directed mutagenesis and molecular modelling, J Biol Chem, **278**: 4250-4257.

Visiers I, Ballesteros JA, Weinstein H. (2002) Three-dimensional representations of G protein-coupled receptor structures and mechanisms, Methods Enzymol, **343**: 329-371.

Xie T, Chen J, Ding DF. (1999) An Evolutionary Trace Method for Functional Prediction of Genomes. Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai), **31**: 433-439.

Zeng FY, Hopp A, Soldner A, Wess J. (1999) Use of a disulfide cross-linking strategy to study muscarinic receptor structure and mechanisms of activation, J Biol Chem, **274**:16629-16640.

Zhu S, Huys I, Dyason K, Verdonck F, Tytgat J. (2004) Evolutionary trace analysis of scorpion toxins specific for K-channels, Proteins, **54**: 361-370.

Chapter 3

Tracing evolutionary pressure

Motivation: Recent advances in sequencing techniques have yielded enormous amounts of protein sequence data from various species. This large dataset allows sequence comparison between paralogous and orthologous proteins to identify motifs or functional positions that account for the differences of functional subgroups ("specificity" positions). Algorithms such as SDPpred and the two-entropies analysis (TEA, as in Chapter 2) have been developed to detect such specificity positions from a multiple sequence alignment (MSA) grouped into classes according to certain biological functions. Other algorithms such as TreeDet compute a classification and then predict specificity positions associated with it. However, there are still many unresolved questions: Was the optimal subdivision of a protein family achieved? Do the definitions at different levels of the phylogenetic tree affect the prediction of specificity positions? Can the whole phylogenetic tree be used instead of only one level in it to predict specificity positions?

Results: In Chapter 3 we present a novel method, TEA-O (Two-Entropies Analysis - Objective), to trace the evolutionary pressure from the root to the branches of the phylogenetic tree. At each level of the tree, a TEA plot is produced to capture the signal of the evolutionary pressure. A consensus TEA-O plot is composed from the whole series of plots to provide a condensed representation. Positions related to functions that evolved early (conserved) or later (specificity) are close to the lower left or upper left corner of the TEA-O plot, respectively. This novel approach allows an unbiased, user-independent, analysis of residue relevance in a protein family.

We compared our TEA-O method with various algorithms using both synthetic and real protein sequences. The results show that our method is robust, sensitive to subtle differences in evolutionary pressure during evolution and comprehensive because all positions in the MSA are presented in the consensus plot.

Available: All computer programs and datasets used in this work are available at http://nava.liacs.nl/kye/TEA-O/ for academic use.

Based upon Kai Ye, Gert Vriend and Adriaan P. IJzerman
Bioinformatics. 24: 908-915 (2008)

Introduction

A crucial aspect in protein sequence analysis is the identification of functional sites such as ligand binding sites, active sites, protein-protein interaction sites, signal sequences, and post-translational modification sites. Traditionally, the residues that are conserved in all members of a protein family are assembled as motifs and correlated to the main function of that protein family. In this way the ATP-binding motif (Walker et al., 1982), the zinc-finger motif (Klug and Rhodes, 1987) and the leucine-zipper motif (Landschulz et al., 1988) were discovered early on. Currently, these functional sites have been collected in databases such as PROSITE (Hulo et al., 2006), Pfam (Bateman et al., 2004), BLOCKS (Henikoff et al., 2000) and PRINTS (Attwood et al., 2003). Recent advances in sequencing techniques have yielded enormous amounts of sequence data from very many species. This prompted a further development of techniques to identify the residues that account for functional differences among subfamilies (del Sol Mesa et al., 2003; Gloor et al., 2005; Kalinina et al., 2004; Kuipers et al., 1997; Mirny and Gelfand, 2002; Oliveira et al., 2002; Pirovano et al., 2006; Chapter 2). We call positions that are conserved in all members of a protein family "conserved positions" and those that are conserved within subfamilies but divergent among them "specificity positions".

The Evolutionary Trace (ET) method recursively searches for the most conserved positions along the phylogenetic tree, from the root to each individual branch. It ranks globally conserved positions high. Positions that show diversity even within a group of highly homologous proteins are placed at the end of the ranking. Recently ET has been applied to zinc binding domains (Lichtarge et al., 1997), steroid receptors (Raviscioni et al., 2006) and G protein-coupled receptors (Madabushi et al., 2004).

Mirny and Gelfand used 'mutual information' to identify positions conserved within orthologs but different among paralogs (Mirny and Gelfand, 2002). The orthologs and paralogs were defined through elaborate sequence comparison. Kalinina et al. incorporated some features of the method proposed by Mirny and Gelfand, and presented their joint method as an SDPpred web server (http://bioinf.fbb.msu.ru/SDPpred) (Kalinina et al., 2004; Mirny and Gelfand, 2002).

While some methods such as SDPpred require a definition of subfamilies in the MSA as input in order to predict specificity positions (Gu, 2006; Kalinina et al., 2004; Mirny and Gelfand, 2002; Pirovano et al., 2006; Ye et al., 2006), many others compute a classification and then search for specificity positions that are associated with the detected classification (del Sol Mesa et al., 2003; Donald and Shakhnovich, 2005; Hannenhalli and Russell, 2000). For instance, in the paper by del Sol Mesa et al. (del Sol Mesa et al., 2003), TreeDet, which contains three algorithms (http://www.pdg.cnb.uam.es/Servers/treedet/), "The Level Entropy Method" (S-method), "Mutational Behavior Method" (MB-method) and "SequenceSpace Automatization Method" (SS-method), was developed to reveal "Tree-determinant residues".

The S-method uses relative entropy to search for an optimal splitting of the input MSA and then considers positions conserved within classes but different among classes as the tree-determinants. The SS-method applies principal component analysis of the multiple sequence alignment to compute an optimal number of clusters and then searches for the positions conserved in, but different among, clusters. The MB-method looks for positions in the MSA whose mutational behavior resembles the phylogeny of that MSA. Different as those methods are in searching for globally conserved or specificity positions, they all aim to derive one scoring function to rank positions. The combination of two scoring functions, however, may yield a greater resolution. Oliveira et al. demonstrated that a scatter plot of variability versus Shannon entropy provided better residue classification than with either variability or Shannon entropy alone (Oliveira et al., 2003).

Building on this principle, we showed that the combination of overall Shannon entropy and the average value of Shannon entropy within subfamilies instead of overall variability lead to a better classification of specificity positions (e.g., the ligand binding site of G protein-coupled receptors) than an entropy-variability plot. Globally conserved positions, specificity positions and others are scattered in the lower left, upper left and upper right corners of the plot, respectively. Powerful as our two-entropies analysis (TEA) appeared to be, it also demanded a subjective definition of subfamilies. The combination of the necessity to have a classification and a need for the use of as many sequences as possible makes the TEA method expensive in terms of human intervention. One solution may be to detect subfamilies for the user as TreeDet does. Since each method has its own mechanism to score for “optimal”, they may classify proteins differently or define subfamilies at different levels of the phylogenetic tree. Hence, a difference in subdivision will be brought to the subsequent prediction of specificity positions as demonstrated in this chapter.

If we consider a subdivision as a single level of the phylogenetic tree, can we then use the entire phylogenetic tree to capture the signal of evolutionary pressure preserved at all levels of the tree? In Chapter 3, we developed the so-called TEA-O (Two-Entropies Analysis - Objective) method to automatically trace the evolutionary pressure along the entire phylogenetic tree to differentiate functional positions including conserved and specificity ones. We compare our approach with the TEA method described in Chapter 2, and with Evolutionary Trace, SDPpred and TreeDet on two protein families, LacI bacterial transcription factors and G protein-coupled receptors.

Materials and methods

TEA-O (Two-Entropies Analysis - Objective)

In Chapter 2, we proposed the two-entropies analysis (TEA) to differentiate between functional positions in a given MSA grouped into multiple subfamilies. Given a definition of subfamilies, we measured the conservation in terms of information entropy for each position

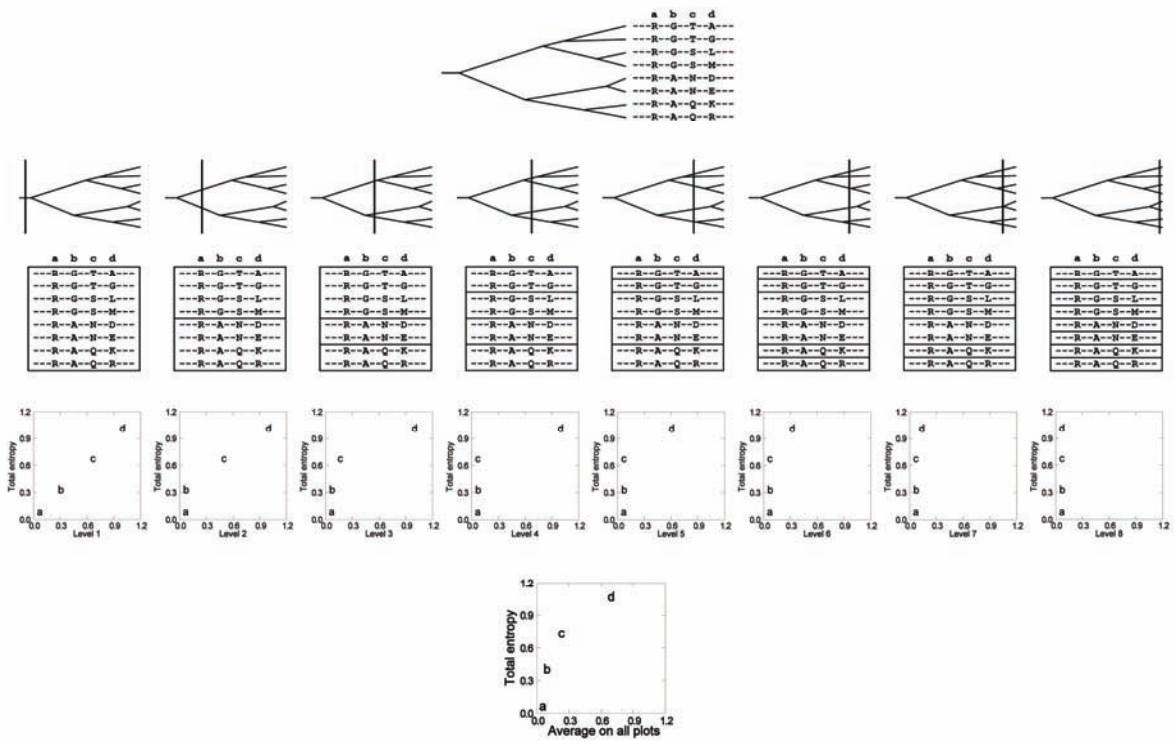


Figure 1. Schematic representation of the TEA-O method. A hypothetical alignment and its phylogenetic tree are used as input. The vertical bars show eight ways to define subfamilies. For a given subdivision, we may plot each position by two types of Shannon entropy. The y-coordinate of the position (total entropy) is calculated using all residues at a given position in the MSA as input. The second entropy (entropy at level X) is obtained by first calculating the entropy in each subfamily and then take the average of these values as x-coordinate of the position. In the consensus TEA-O plot, the x coordinates are averaged again among all plots while the y-coordinate, which is independent on subdivisions, remains the same.

at both the entire superfamily and subfamily level. Then we used these two entropy values to produce a plot. In this chapter, we propose a radically different approach to trace the evolutionary pressure from the root to the branches of a phylogenetic tree. To distinguish our new approach from the old one, two-entropies analysis (TEA), we call it the TEA-O (Two-Entropies Analysis - Objective) method.

Figure 1 depicts the work flow of TEA-O. The input for the TEA-O method is a single MSA without a definition of subfamilies. We used p-distance and UPGMA algorithm implemented in MEGA3.1 (Kumar et al., 2004) to construct a phylogenetic tree. As shown in Figure 1, the vertical bar indicates a certain evolutionary distance (often expressed as a percentage sequence similarity) that could be used as a cutoff for calling a set of sequences a subfamily. For example, in the fifth way to define subfamilies, five subfamilies would result: (1), (2), (3,4), (5,6), (7,8). Intuitively, one would like to combine sequence (1) and (2) in one subfamily, but it is equally defendable to move the vertical bar a short distance to the right to

split the subfamily (7, 8). Clearly the definition of subfamilies using the method illustrated in figure 1 is arbitrary and large differences in the locations of residues in the whole series of two-entropies plots are observed if different subfamily definitions are used. In this chapter, we show that the arbitrary nature of the subfamily definition can be overcome by repeating the two-entropies plot generation process for each of the N possible subfamily definitions (N being the number of sequences in the MSA), and using the average of all resulting two-entropies plots.

In addition, we modified the equation of Shannon entropy in order to consider gaps in the alignment as well as the potential unbalance in the sizes of subfamilies. A seemingly easy way to treat gaps in the calculation of information entropy is to consider a gap as the 21st residue as is done in the Evolutionary Trace (ET) method. However, this will lead to the obvious error that the more gaps in one position the more conserved the position is. For example, given an alignment with 100 protein sequences, ET defines a position with 1Y and 99 "-" as more conserved, thus more important, than another one with 98Y and 2 "-". Thus we treated each gap as a different residue in calculating entropy values. For example, the first gap in the position is the 21st residue; the second the 22nd and so on. In this way, positions with a high percentage of gaps have high entropy values.

Since the number of proteins in one group determines the maximum information entropy value, we also adjusted the entropy value according to the size of each group while calculating entropy values as suggested by Valdar (2002).

For a given definition of subfamilies, the average entropy value among all subfamilies at position *i* is given by

$$E_i = \frac{1}{m}\sum_{j=1}^{m} W^j * \left(-\sum_{a=1}^{20}\left(\frac{A_{ia}^j}{n^j} * \ln\frac{A_{ia}^j}{n^j}\right) - G_i^j * \frac{1}{n^j} * \ln\frac{1}{n^j}\right)$$

$$W^j = \frac{1}{\ln S} \text{ (if } n^j < 20,\ S = n^j;\ \text{else } S = 20)$$

in which there are m subfamilies at the level *i* and n^j stands for the number of the proteins (*n*) in a given subfamily *j*. A_{ia}^j is the number of occurrences of one particular amino acid type *a* in a given subfamily *j* at position *i*. G_i^j stands for the number of gaps in a given subfamily *j* at position *i*.

Finally in the consensus plot, the x-coordinate of each position is the average of all TEA plots while the y-coordinate remains the same.

The weighting factor scales the entropy to range [0, 1] so that scoring functions may be easily derived to rank conserved and specificity positions from the consensus TEA-O plot. The x-coordinates of all positions in the TEA-O plot are then standardized to [0, 1]. Since globally

conserved positions are close to the lower left corner of the plot, the scoring function for them is defined as the distance to (0, 0):

$$f_{conserved} = \sqrt{E_x^2 + E_y^2}$$

On the other hand, the specificity positions are close to the upper left corner of the plot so that the scoring function for them is defined as the distance to (0, 1):

$$f_{specificity} = \sqrt{E_x^2 + (E_y - 1)^2}$$

E_x and E_y are the coordinates of a given position in the TEA-O plot.

Datasets for validation

Synthetic dataset In order to demonstrate TEA-O's features and compare it with other methods, we first simulated an MSA according to our simple protein evolution model described below. Note that this model does not aim to include all aspect of protein evolution but to shed light only on how different definitions of subfamilies affect the prediction of specificity positions. For example, we did not introduce any substitution matrices to the model since all methods in this chapter use either variability- or entropy-related scores.

We simulated four generations of protein evolution by modifying the evolutionary pressure between generations. As a start we randomly generated a protein with 200 residues. Then an evolutionary pressure (EP, real number between 0 to 1) was randomly assigned to each of the 200 residues. If a position is assigned an EP close to 1, the probability of this residue mutating to another one will be small, and this position will be conserved during evolution; if the EP is close to 0, the chance of mutation will be high, and this position will be divergent. We marked those positions of the protein with an evolutionary pressure bigger than 0.85 as the globally conserved ones and called them EP1. Subsequently we multiplied the ancestor protein 5 times to get the second generation of proteins and mutated each position in every protein according to the assigned EP as follows. A random number between 0 and 1 was generated, whereafter it was compared with its EP. If the generated number was larger than EP, one of the 20 amino acid subtypes was randomly chosen to replace the current residue. In this case, we obtained 5 proteins in the second generation.

Proteins obtain new functional sites during evolution. If certain advantages are associated with these new functional sites, the evolutionary pressures will be tightened on the positions involved. Hence we modified the EP for some positions after we obtained each generation in order to simulate such gain of function during evolution. We randomly chose 20 positions (10% of the sequence length) with EP smaller than 0.5 and randomly modified their EP to a value between 0.85 and 1.0. We called these 20 positions EP2. After that we multiplied each

member of the second generation 5 times and mutated them according to the updated evolutionary pressure. This yielded the third generation with 25 proteins.

We repeated such modification of EP, multiplication and mutation to get the fourth generation of proteins as well as a list of EP3 positions. In this way we obtained 1*5*5*5=125 proteins as the fourth generation of proteins.

Real protein sequences

LacI bacterial transcription factors The LacI family is one of the largest families of bacterial transcription factors. Dr. Mirny (MIT, USA) kindly provided us with the MSA of this family and the definition of subfamilies. The multiple sequence alignment of 55 LacI proteins was classified into 15 subfamilies (Mirny and Gelfand, 2002). Extensive experimental (Glasfeld et al., 1999; Lehming et al., 1990; Sartorius et al., 1991; Suckow et al., 1996) and structure information (Bell and Lewis, 2001; Schumacher et al., 1997) allowed us to evaluate our prediction. Suckow et al replaced position 2 to 329 of the Lac repressor with 12 or 13 of the 20 naturally occurring amino acids. These 4000 well-defined mutants yielded a functional classification for each position. The non-conserved residues in group XI (IPTG contacts) were defined as the specificity positions. The specificity positions based on experimental evidence were L73, N125, P127, A145, S191, S193, W220, Q248, T276, and F293. We also defined specificity positions based on structural information. The non-conserved residues within 4 Å distance to the ligand, ONPF (ortho-nitrophenyl-beta-D-fucopyranoside), in its crystal structure (PDB: 1JWL), were defined as specificity positions; these were L73, S191, V192 and W220.

G protein-coupled receptors (GPCRs) GPCRs are integral cell membrane proteins involved in signal transduction. Such a mediatory role makes them important drug targets. We extracted the MSA of class A GPCRs from the latest version of the GPCRDB (June 2006 release 10.0, http://www.gpcr.org/7tm/) (Horn et al., 2003). This yielded an MSA of 2065 protein sequences with an average sequence identity of 25.8%. The MSA was classified into 77 subfamilies according to the recognition of endogenous ligands since SDPpred and TEA require such information in predicting specificity positions.

The specificity positions based on structural information were defined as residues within 4 Å distance of the endogenous ligand retinal but not being a member of the well-known conserved aromatic cluster in helix 6 (Javitch et al., 1998). They are E113, A117, T118, G121, E122, M207, A292 and K296 of bovine rhodopsin. We also defined specificity positions based on experimental evidence reviewed by Shi and Javitch (2002). Positions were considered to be part of the specificity recognition site if they were located in a TM region and implicated in ligand binding in aminergic receptors. The specificity positions based on experimental evidence are T94, T97, E113, G114, A 117, T118, G121, C167, L172, F203,

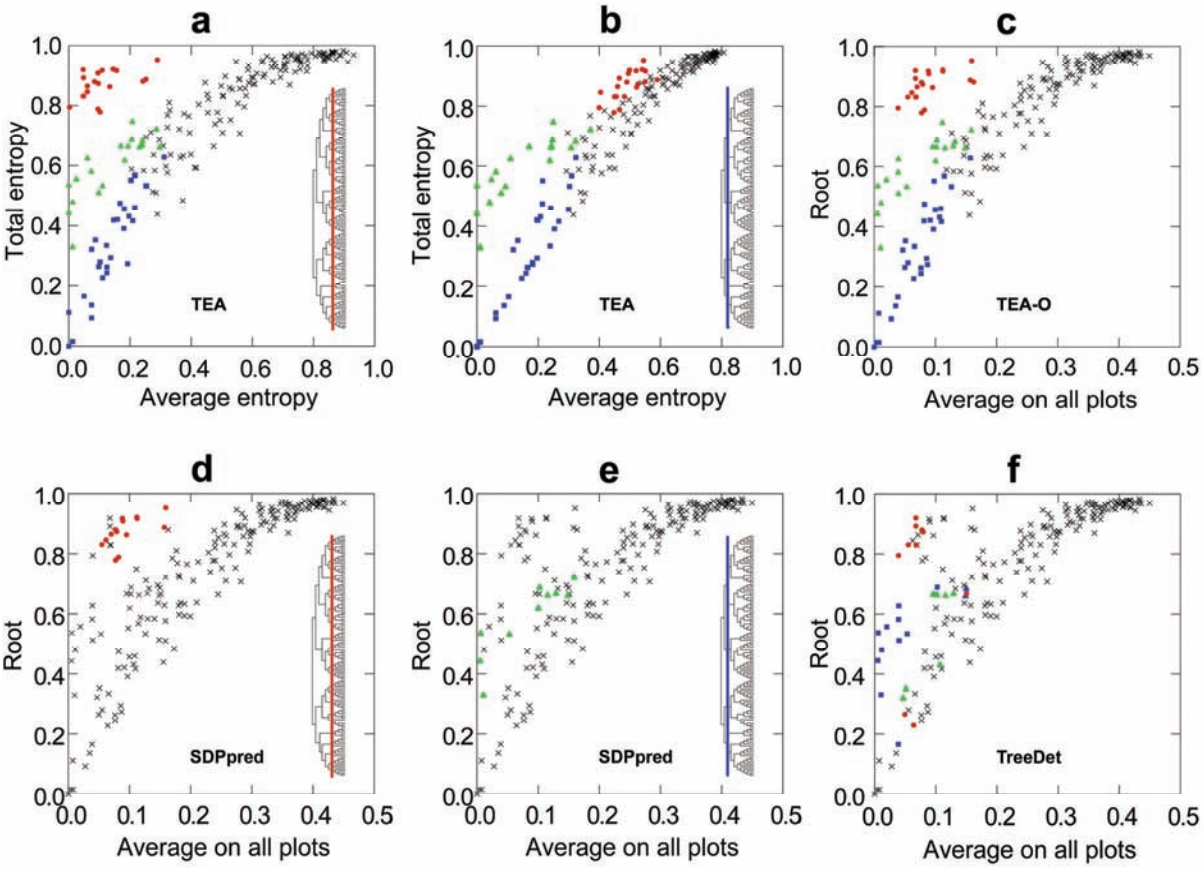


Figure 2. Comparison among TEA (a, b), TEA-O (c), SDPpred (d, e) and TreeDet (f) in the identification of three groups of functional positions (EP1, EP2 and EP3) using the synthetic dataset as an input. a) TEA plot using the MSA grouped at the third generation level. ■EP1; ▲EP2; ●EP3; b) TEA plot using the MSA grouped at the second generation level. ■EP1; ▲EP2; ●EP3; c) TEA-O plot. ■EP1; ▲EP2; ●EP3; d) and e) Prediction of specificity positions by SDPpred were mapped on the TEA-O plot. ▲specificity positions predicted by SDPpred using the MSA grouped at the second generation level; ●specificity positions predicted by SDPpred using the MSA grouped at the third generation level; f) Specificity positions predicted by TreeDet (SS-method and MB-method) were mapped on the TEA-O plot. ■predicted as specificity positions by both SS-method and MB-method; ▲predicted only by the MB-method; ●predicted only by the SS-method.

V204, M207, F208, H211, A272, A292, F293 and K296 according to their numbering in bovine rhodopsin.

Results

Specificity positions in the synthetic dataset

Our new TEA-O method does not require the user to define subfamilies within a protein family while many methods including TEA in Chapter 2 demand such a definition. Thus, we first examined how different definitions of subfamilies affect the identification of specificity positions using the synthetic dataset. When we divided the MSA at the third generation level,

we obtained 25 subfamilies, each containing 5 proteins. As shown in Figure 2a, when we used such a definition, our previous method, TEA identified EP1 (conserved), EP2 and EP3 (specificity). The separation of EP3 from other positions is much better than of EP2. However, if we defined subfamilies at the second generation level (5 subfamilies, each of them contains 25 proteins), separation of EP2 remained the same but EP3 mixed with other less determining positions (Figure 2b).

When we used the new TEA-O method to analyze the same MSA without a definition of subfamilies, we obtained a series of TEA plots when we traced the evolutionary pressure from the root to the leaves of the phylogenetic tree. The positions under different evolutionary pressures (EP1, EP2, EP3 and others) migrate differently as shown in the animation in the supporting material. When we averaged the entire series of plots, we obtained a TEA-O plot (Figure 2c), which is similar to the plot generated by the two-entropies analysis using the MSA grouped at the third generation (Figure 2a). Thus the new TEA-O method maintains the same separation ability as the previous TEA method even when there is less information in the input. From Figure 2a and 2b we learned that definitions of subfamilies at different levels of the phylogenetic tree dramatically change the performance of the TEA method. We next examined SDPpred using the same MSA with the subfamily definition at the third or the second generation. We mapped the predictions of SDPpred on our TEA-O plot (Figure 2c) as shown in Figure 2d and 2e. When the definitions of subfamilies are at the second or the third generation, SPDpred identified part of EP2 or EP3, respectively. This clearly demonstrates that grouping of protein sequences has a profound impact on the prediction of specificity positions for algorithms that require such a definition of subfamilies.

To compare the prediction of TEA-O on the synthetic dataset with TreeDet which detects subfamilies by itself, we directly mapped the prediction of the latter method on the TEA-O plot. TreeDet contains three different algorithms: the MB-method, the SS and the S-method. The S-method failed to identify any specificity positions. As shown in Figure 2f, the MB-method and SS-method behaved very differently since the MB-method failed to identify any of EP3 but found more EP2 than the SS-method. Both the MB and SS-methods mislabeled a few globally conserved positions (EP1) as specificity positions.

Prediction of specificity positions of real protein sequences

The performance of TEA-O was compared with the SDPpred, TEA, TreeDet and Evolutionary Trace methods in predicting specificity positions of both LacI bacterial transcription factors and G protein-coupled receptors using either experimental or structural information (Figure 3). TEA-O performs comparably to or even better than SDPpred and TEA and, importantly, requires no classification information from the user. The poor performance of TreeDet may be due to misclassification. If we take the manual classification of 15 subfamilies in the LacI protein family by Mirny and Gelfand for comparison, TreeDet detected only 6 of these and some proteins from the same subfamily were identified in the

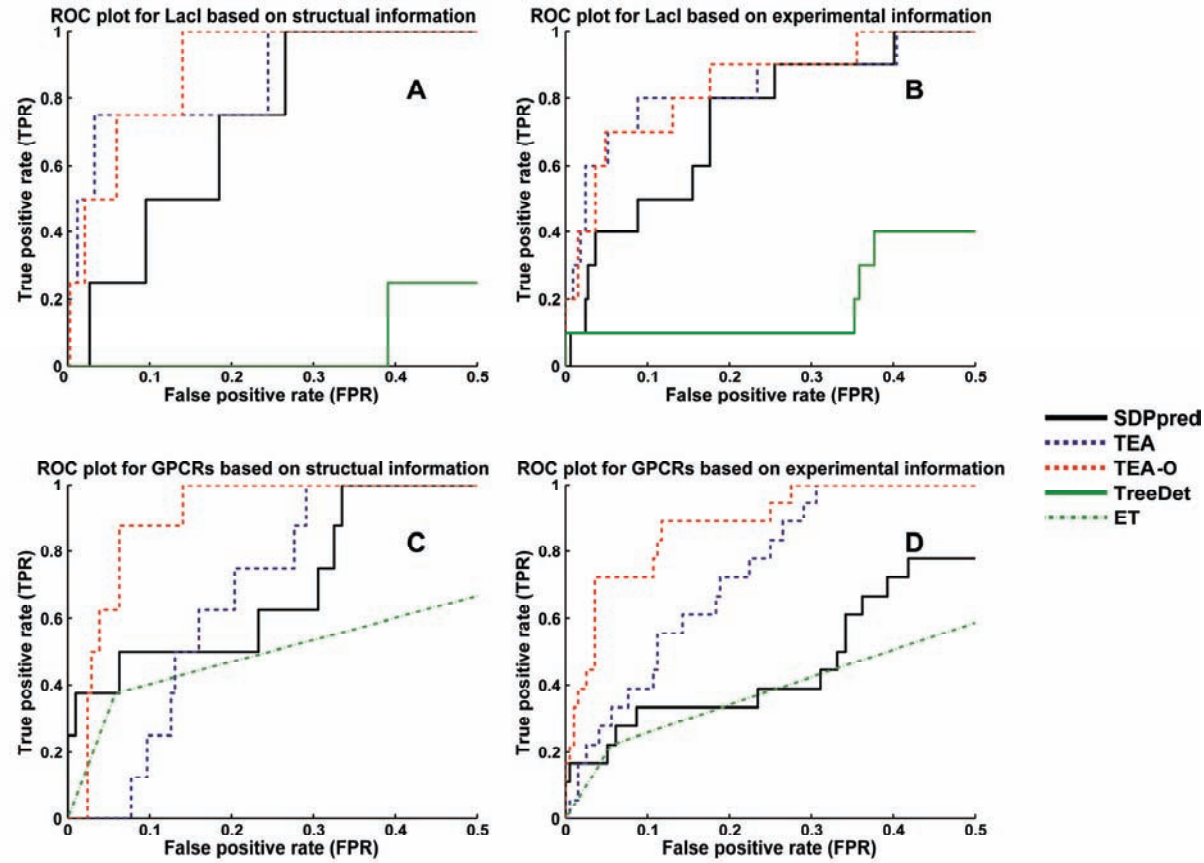


Figure 3. ROC plots for both LacI and GPCRs to evaluate specificity prediction by SDPpred, TEA, TEA-O, TreeDet (only LacI, see text) and evolutionary trace (ET) based on both structural and experimental information. A) the ROC plot for the LacI family based on structural information; B) the ROC plot for the LacI family based on experimental information; C) the ROC plot for GPCRs based on structural information; D) the ROC plot for GPCRs based on experimental information.

wrong subfamilies (see supporting material for more information on the differences between manual and automatic classifications in the LacI family).

We could not apply TreeDet on the GPCRs dataset because the number of proteins in this family (2065) is too large for TreeDet (maximally 200). Evolutionary Trace did not identify specificity positions of GPCRs and LacI with either of the two implementations of this method. The first implementation of evolutionary trace, “ET report maker”, accepts a single protein but not an MSA as input. The second implementation, ET viewer, does accept an MSA but it ranks positions according to their importance with respect to conservation. In the list, the conserved positions are ranked high while it is not stated which part of the list is related to the specificity positions. Madabushi et al. used ET to compute the most conserved residues in all GPCRs as well as in rhodopsins (a particular subfamily of GPCRs). Then they defined the rhodopsin-specific positions by subtracting conserved residues in all GPCRs from those conserved in rhodopsin (Madabushi et al., 2004). As shown in Figure 3 the performance

of ET on the GPCRs dataset is comparable to SDPpred, TEA and TEA-O, when based on structure information but performs the worst when based on experimental evidence (see Materials and Methods for definitions).

Discussion

Comparison with other methods

It is generally believed that the conserved positions in an MSA are functionally important. Evolutionary Trace (ET) is one of the first methods that systematically explored the correlation between residue conservation and functional importance. More importantly, the authors of the method proposed that specificity positions are conserved within subfamilies but divergent among them. However, both implementations of ET mainly focus on conserved positions and no automatic mechanism has been provided to identify specificity positions. For example, one may use ET viewer to find conserved positions in the entire MSA as well as in one particular subfamily. Then the user has to manually subtract the positions conserved in the entire MSA from those conserved in the subfamily. This yields a group of residues conserved in the subfamily only, but without ranking.

We believe that specificity positions in the same protein family should be largely overlaid while each subfamily may slightly modify its specificity site to accommodate the difference of the modulator. The way ET viewer is implemented to identify specificity positions in one particular subfamily overestimates the difference among the specificity sites of various subfamilies. Thus in practice, ET ignores the conservation signal from the peer subfamilies to eliminate false positives while tracing the specificity positions in one subfamily. Many "conserved" positions identified by ET in one subfamily may not be functionally important but just the consequence of a small population of proteins or a short evolution history since the subfamily members began to evolve.

Although TreeDet does not require a definition of subfamilies by the user, it tries to find an "optimal" definition of subfamilies early in the calculation. As we demonstrated in Figure 2, any prediction based on only one level of the phylogenetic tree will introduce bias to positions associated with that level of the tree. In addition, misclassification will jeopardize the prediction (as shown in Figures 3A and 3B).

Figure 4. The endogenous ligands of dopamine and adrenergic receptors. The compound noradrenaline prefers alpha-adrenergic receptors, whereas adrenaline prefers beta-adrenergic receptors.

The new TEA-O is better than TEA (Figure 3) mainly because its prediction is not biased to a particular classification. When we produce a classification either manually or computationally, we are speculating the “optimal” cutoff to recognize certain proteins with similar biological functions. In the case of GPCRs, we defined subfamilies based on the recognition of endogenous ligands according to the web-publication (http://www.iuphar-db.org/GPCR/) of the International Union of Basic and Clinical Pharmacology (IUPHAR). Even in this international standard about GPCRs classification, the definition of subfamilies is quite arbitrary. For example, dopamine receptors and adrenergic receptors are considered as two subfamilies by IUPHAR. The adrenergic receptors are further classified as alpha-adrenergic and beta-adrenergic receptors. The compound noradrenaline prefers alpha-adrenergic receptors, whereas adrenaline prefers beta-adrenergic receptors. As shown in Figure 4, the only difference between the two compounds is a methyl group on the terminal amino group. In addition, the compound dopamine lacks the beta-hydroxy group of noradrenaline and receptors that recognize dopamine are defined as one family, dopamine receptors. If we look at these three molecules, it is indeed arbitrary to group alpha- and beta-adrenergic receptors as one family and dopamine receptor as another. One might even define all of them as one family if we consider the differences among these three compounds negligible. As we demonstrated in Figures 2a, 2b, 2d and 2d, any classification will introduce a bias to the positions associated strongly with it. The classification of GPCRs we used for the TEA method is just one of the possible classifications. It has been also shown in the LacI dataset that when classification is sufficiently optimal, TEA-O performs as well as TEA. This indicates that by averaging the entire series of TEA pots, the specificity signal in TEA-O plot is as strong as in the TEA plot when optimal classification is given.

Further development of TEA-O

In this chapter, we used information entropy to measure conservation for a given position in the alignment. Information entropy as a measure is certainly more informative than variability since it considers not only the number of amino acid types but also their relative frequencies. However, information entropy treats amino acids as 20 independent symbols and ignores the fact that amino acids can be quite similar. For two positions in an alignment, e.g., one with 50% D and 50% E and the other with 50% D and 50% W, information entropy as a measure yields the same score, although the former combination is more conserved as both D and E are negatively charged. Ideally, a conservation measure should also account for the mutation type.

We introduced a weighing factor scaling the entropy values in between 0 and 1. Such a weighing factor is only related to the number of sequences in a given alignment, and does not account for sequence similarity between various alignments. Suppose we have two subfamilies, A and B, for which the sequences in A are very similar to each other while those in B are very different. If the same substitution pattern is observed in both A and B, a

conservation measure should yield different scores in the two subfamilies for such a position. Today, most specificity prediction algorithms still use relatively simple conservation measures such as variability and information entropy. In a future study we aim to use a more sophisticated conservation measure such as Rate4Site (Landau et al., 2005; Pupko et al., 2002) instead of the information entropy we used in the present algorithm.

Conclusion

In Chapter 3 we present a novel method coined TEA-O (Two-Entropies Analysis - Objective) to relate residues to their potential functions from an MSA. Compared to TEA, TEA-O requires only an MSA to predict both conserved and specificity positions. This novel approach allows an unbiased, user-independent, analysis of residue relevance in a protein family.

TEA-O traces the evolutionary pressure from the root to the branches of the phylogenetic tree. At each level of the tree, a two-entropies plot is produced to capture the signal of the evolutionary pressure. The whole series of plots comprehensively reproduces the evolutionary pressure on each position in every period of evolutionary history. A consensus plot is composed from the whole series of plots to provide a condensed representation. Positions related to the functions that evolved early (conserved) are close to the lower left corner of the consensus plot while those related to functions evolved later (specificity) are located close to the upper left corner of the plot. Scoring functions for ranking conserved and specificity positions are provided.

We compared our TEA-O method with various algorithms using both synthetic and real protein sequences. The results show that our method is robust, sensitive to subtle differences in evolutionary pressure during evolution and comprehensive since all positions in the MSA are presented in the consensus plot.

References

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C. (2003) PRINTS and its automatic supplement, prePRINTS, *Nucleic Acids Res*, **31**, 400-402.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database, *Nucleic Acids Res*, **32**, D138-141.

Bell, C.E. and Lewis, M. (2001) Crystallographic analysis of Lac repressor bound to natural operator O1, *J Mol Biol*, **312**, 921-926.

del Sol Mesa, A., Pazos, F. and Valencia, A. (2003) Automatic methods for predicting functionally important residues, *J Mol Biol*, **326**, 1289-1302.

Donald, J.E. and Shakhnovich, E.I. (2005) Determining functional specificity from protein sequences, *Bioinformatics*, **21**, 2629-2635.

Donald, J.E. and Shakhnovich, E.I. (2005) Predicting specificity-determining residues in two large eukaryotic transcription factor families, *Nucleic Acids Res*, **33**, 4455-4465.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information, *Bioinformatics*, **19**, 163-164.

Glasfeld, A., Koehler, A.N., Schumacher, M.A. and Brennan, R.G. (1999) The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions, *J Mol Biol*, **291**, 347-361.

Gloor, G.B., Martin, L.C., Wahl, L.M. and Dunn, S.D. (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions, *Biochemistry*, **44**, 7156-7165.

Gu, X. (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences, *Mol Biol Evol*, **23**, 1937-1945.

Hannenhalli, S.S. and Russell, R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments, *J Mol Biol*, **303**, 61-76.

Henikoff, J.G., Greene, E.A., Pietrokovski, S. and Henikoff, S. (2000) Increased coverage of protein families with the blocks database servers, *Nucleic Acids Res*, **28**, 228-230.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F.E. and Vriend, G. (2003) GPCRDB information system for G protein-coupled receptors, *Nucleic Acids Res*, **31**, 294-297.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. and Sigrist, C.J. (2006) The PROSITE database, *Nucleic Acids Res*, **34**, D227-230.

Javitch, J.A., Ballesteros, J.A., Weinstein, H. and Chen, J. (1998) A cluster of aromatic residues in the sixth membrane-spanning segment of the dopamine D2 receptor is accessible in the binding-site crevice, *Biochemistry*, **37**, 998-1006.

Kalinina, O.V., Mironov, A.A., Gelfand, M.S. and Rakhmaninova, A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families, *Protein Sci*, **13**, 443-456.

Kalinina, O.V., Novichkov, P.S., Mironov, A.A., Gelfand, M.S. and Rakhmaninova, A.B. (2004) SDPpred: A tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins, *Nucleic Acids Res*, **32**, W424-428.

Klug, A. and Rhodes, D. (1987) Zinc fingers: A novel protein fold for nucleic acid recognition, *Cold Spring Harb Symp Quant Biol*, **52**, 473-482.

Kuipers, W., Oliveira, L., Vriend, G. and Ijzerman, A.P. (1997) Identification of class-determining residues in G protein-coupled receptors by sequence analysis, *Receptors Channels*, **5**, 159-174.

Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, *Brief Bioinform*, **5**, 150-163.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. and Ben-Tal, N. (2005) ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures, *Nucleic Acids Res*, **33**, W299-302.

Landschulz, W.H., Johnson, P.F. and McKnight, S.L. (1988) The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins, *Science*, **240**, 1759-1764.

Lehming, N., Sartorius, J., Kisters-Woike, B., von Wilcken-Bergmann, B. and Muller-Hill, B. (1990) Mutant lac repressors with new specificities hint at rules for protein-DNA recognition, *Embo J*, **9**, 615-621.

Lichtarge, O., Yamamoto, K.R. and Cohen, F.E. (1997) Identification of functional surfaces of the zinc binding domains of intracellular receptors, *J Mol Biol*, **274**, 325-337.

Madabushi, S., Gross, A.K., Philippi, A., Meng, E.C., Wensel, T.G. and Lichtarge, O. (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions, *J Biol Chem*, **279**, 8126-8132.

Mirny, L.A. and Gelfand, M.S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors, *J Mol Biol*, **321**, 7-20.

Mirny, L.A. and Gelfand, M.S. (2002) Using orthologous and paralogous proteins to identify specificity determining residues, *Genome Biol*, **3**, PREPRINT0002.

Oliveira, L., Paiva, A.C. and Vriend, G. (2002) Correlated mutation analyses on very large sequence families, *Chembiochem*, **3**, 1010-1017.

Oliveira, L., Paiva, P.B., Paiva, A.C. and Vriend, G. (2003) Identification of functionally conserved residues with the use of entropy-variability plots, *Proteins*, **52**, 544-552.

Pirovano, W., Feenstra, K.A. and Heringa, J. (2006) Sequence comparison by sequence harmony identifies subtype-specific functional sites, *Nucleic Acids Res*, **34**, 6540-6548.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. and Ben-Tal, N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, *Bioinformatics*, **18** Suppl 1, S71-77.

Raviscioni, M., He, Q., Salicru, E.M., Smith, C.L. and Lichtarge, O. (2006) Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors, *Proteins*, **64**, 1046-1057.

Sartorius, J., Lehming, N., Kisters-Woike, B., von Wilcken-Bergmann, B. and Muller-Hill, B. (1991) The roles of residues 5 and 9 of the recognition helix of Lac repressor in lac operator binding, *J Mol Biol*, **218**, 313-321.

Schumacher, M.A., Glasfeld, A., Zalkin, H. and Brennan, R.G. (1997) The X-ray structure of the PurR-guanine-purF operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated hydrogen bond to corepressor specificity and binding affinity, *J Biol Chem*, **272**, 22648-22653.

Shi, L. and Javitch, J.A. (2002) The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop, *Annu Rev Pharmacol Toxicol*, **42**, 437-467.

Suckow, J., Markiewicz, P., Kleina, L.G., Miller, J., Kisters-Woike, B. and Muller-Hill, B. (1996) Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure, *J Mol Biol*, **261**, 509-523.

Valdar, W.S. (2002) Scoring residue conservation, *Proteins*, **48**, 227-241.

Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold, *Embo J*, **1**, 945-951.

Ye, K., Lameijer, E.W., Beukers, M.W. and IJzerman, A.P. (2006) A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors, *Proteins*, **63**, 1018-1030.

Chapter 4

Multi-RELIEF: A method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting

Motivation: Identification of residues that account for protein function specificity is crucial, not only for understanding the nature of functional specificity, but also for protein engineering experiments aimed at switching the specificity of an enzyme, regulator or transporter. Available algorithms generally use multiple sequence alignments to identify residue positions conserved within subfamilies but divergent in between. However, many biological examples show a much subtler picture than simple intra-group conservation versus inter-group divergence.

Results: We present multi-RELIEF, a novel approach for identifying specificity residues that is based on RELIEF, a state-of-the-art Machine-Learning technique for feature weighting. It estimates the expected "local" functional specificity of residues from an alignment divided in multiple classes. Optionally, 3D structure information is exploited by increasing the weight of residues that have high-weight neighbours. Using ROC curves over a large body of experimental reference data, we show that (a) multi-RELIEF identifies specificity residues for the seven test sets used, (b) incorporating structural information improves prediction for specificity of interaction with small molecules and (c) comparison of multi-RELIEF with four other state-of-the-art algorithms indicates its robustness and best overall performance.

Based upon Kai Ye, K. Anton Feenstra, Jaap Heringa, Adriaan P. IJzerman and Elena Marchiori. Bioinformatics 2008, 24(1):18–25
A web-server implementation of multi-RELIEF is available at www.ibi.vu.nl/programs/multirelief.

Introduction

Many homologous protein families have a common biological function but different specificity towards substrates, ligands, effectors, proteins and other interacting molecules. All these interactions require certain specificity. Identifying crucial residues for this specificity is a prerequisite for understanding the nature of functional specificity, for planning experiments on functional analysis or protein redesign, and for guiding point mutations aimed at switching the specificity of an enzyme, regulator or transporter.

In order to detect specificity residues, advanced computational techniques are needed, because of a great variety of functional specificities observed in nature and the vast amount of protein sequence data. A number of algorithms have been proposed in recent years for detecting specificity residues from a multiple sequence alignment (MSA) (Bickel et al., 2002; Carro et al., 2006; Del Sol Mesa et al., 2003; Feenstra et al., 2007; Gu, 2006; Hannenhalli and Russell, 2000; Kalinina et al., 2004; Mihalek et al., 2004; Ye et al., 2006 as in Chapter 2). Most algorithms employ information-entropy-related scoring functions (Shenkin et al., 1991) to rank residue positions according to the association with the subfamilies (Whisstock and Lesk, 2003). While many algorithms require a predefined subdivision of the MSA into classes, some induce a grouping on the fly.

The SDPpred method (Kalinina et al., 2004) uses mutual information to identify residue positions in which amino acid distributions correlate with the subfamily grouping (Mirny and Gelfand, 2002).

The Two-Entropies Analysis algorithm (TEA, Chapter 2) creates a 2D plot of residue conservation in terms of Shannon entropy at both superfamily and subfamily level. Functional sites such as conserved or specificity residues can be distinguished easily from other residues.

The TreeDet approach introduced by Del Sol Mesa et al. (2003) contains three algorithms for detecting so-called tree-determinant residues from an unpartitioned MSA. The Level Entropy (S) method first uses relative entropy to search for an optimal grouping of the alignment and then considers positions conserved within classes but different among classes as the tree-determinants. The Sequence Space Automatization (SS) method applies principle component analysis to the alignment and computes an optimal number of clusters and the residues that correspond to them. Finally, the Mutational Behavior (MB) method looks for residues whose mutational behavior resembles the phylogeny of the alignment.

The Sequence Harmony (SH) method (Feenstra et al., 2007; Pirovano et al., 2006) scores compositional overlap between two user-specified groups. The algorithm does not exploit the notion of subfamily conservation but focuses on compositional differences between the subfamilies.

In this chapter, we introduce multi-RELIEF, a new algorithm for identifying specificity residues from a given MSA and predefined multiple classes using “local” conservation

properties. The approach is based on a state-of-the-art Machine-Learning technique for feature weighting, called RELIEF, which exploits the notion of locality for estimating relevance of attributes in discriminating samples from two classes (Kononenko, 1994). In the biological context considered here, locality corresponds to *sequence space* (Landgraf et al., 2001).

Multi-RELIEF estimates the expected "local" specificity of residues, by comparing each sequence with the most similar sequence in the same class and with the most similar in opposite classes. The nearest neighbor sequences are selected based on global identity. A residue is considered relevant if it has high local specificity with respect to at least one pair of classes.

While other algorithms consider residue positions independently, multi-RELIEF considers global sequence similarity while scoring each residue. Furthermore, the method can cope with sub-family classifications derived from phylogeny, which generally are heterogeneous. Misclassification, a general error that can arise from, e.g., misannotation, will result in a close match between some opposing classes. Multi-RELIEF is able to "recover" the innate specificity of a class, whenever one of the other classes can be contrasted to it. This alleviates the problem of downweighting the relevance of residue positions, e.g., in cases where a single class is "polluted" with a misplaced sequence.

Multi-RELIEF can optionally include 3D structural information, if available. It does this by employing a new heuristic based on the assumption that a specificity residue does not evolve in isolation, but within a functional cluster in the protein structure. This means that a residue would be more likely to be a specificity residue if its neighboring residues are also specific.

To test our novel approach thoroughly, seven experimentally determined benchmark sets were considered, taken from five widely studied protein families: G protein-coupled receptors (GPCRs), the LacI family of bacterial transcription factor, the Ras-superfamily of small GTP-ases, the MIP-family of integral membrane transporters and the Smad family of transcription factors. The performance of multi-RELIEF was compared with TEA and SDPpred (both acting on multiple classes), TreeDet/MB (no class division required) and SH (acting on two classes). Using ROC curves we show that (a) multi-RELIEF identifies specificity residues, (b) incorporating structural information improves prediction for specificity of interaction with small molecules and (c) comparison of multi-RELIEF with other algorithms indicates its robustness and best overall performance.

Methods

RELIEF

Many interesting feature weighting algorithms based on different approaches have been introduced in Machine Learning (Guyon and Elisseeff, 2003). One particular class uses a

multivariate “filter” prior to the construction of a model (the classifier) to quantify the relevance of features as to their ability to jointly discriminate between classes.

RELIEF is considered one of the most successful filter multivariate feature weighting algorithms (Guyon and Elisseeff, 2003), due to its simplicity and effectiveness (Kononenko, 1994). We recently applied RELIEF for selecting specificity residues (“subtype specific functional sites”) from protein sequences of the Smad receptor binding family (Marchiori et al., 2006).

Given samples from two classes, RELIEF iteratively assigns weights to features based on how well they separate samples from their nearest neighbor (*nnb*) within the same class relative to that within the opposite class (Marchiori et al., 2006). To do this, RELIEF employs a feature weight vector. At each iteration, one sequence *seq* is selected. The weights are updated by adding the “difference” between *seq* and its *nnb* from the opposite class, *miss*(*seq*), and subtracting the difference between *seq* and its nnb from the same class, *hit*(*seq*). We define *nnb* for a sequence *seq* to class *l* as $nnb(seq)=argmin\ \{d(seq, x), x \in X_l\}$ where *d* denotes the Hamming distance between strings [e.g., *d*(“ALM”, “VLM”)=1]. The difference between two sequences *seq1_seq2* is a vector representing matches (0) and mismatches (1) between residues (e.g. “ALM”-”VLM”=100). This procedure is iterated over all sequences of the dataset. The computational complexity of RELIEF is $O((number\ of\ sequences)^2 * (number\ of\ positions))$.

A residue position (“site”) will obtain best weight if it has maximal “local” specificity over all triplets of a sequence, its nearest neighbor in the same, and that in the opposite class, i.e., local in *sequence space*. Thus if a residue position is conserved within each class but divergent between classes, then its RELIEF weight will be high. Completely conserved positions and overall divergent positions will get zero weight, while positions that are divergent within subfamilies but conserved between subfamilies will get negative weight.

Multi-RELIEF

RELIEF is a two-class feature weighting algorithm. However, large protein families with a variety of specificities require algorithms acting on multiple classes. Extensions of RELIEF to handle multiple classes have been proposed (Kononenko, 1994; Robnik-Sikonja and Kononenko, 2003; Sun and Li, 2006). For instance, Kononenko (1994) introduced RELIEF-F where the weight vector is updated by the sum of *miss*(*seq*) weighted by the estimated *a priori* probabilities of the classes. Here, we present a new ensemble approach based on random sub-sampling of pairs of classes. The multi-RELIEF algorithm is illustrated below in pseudo-code.

```
Multi-RELIEF
%input: X1,..,Xm (m classes of aligned proteins)
%parameters: nr_iter, nr_sample
%output: multi_W (weights assigned to positions)
nr_positions=total number of positions;
weights=zero vector of size nr_positions;
for i=1: nr_iter
   select randomly two classes
   X=select randomly nr_sample sequences from each selected class
   W_i=apply RELIEF to X
end;
for s=1: nr_positions
   multi_W(s)=(average across positive W_i(s)'s);
end;
return multi_W
```

In multi-RELIEF, multiple runs (*nr_ iter*) of RELIEF are performed. At each run *i*, first two classes are randomly selected. Next, *nr_sample* sequences from each class are randomly selected. Finally, RELIEF is applied to the resulting two classes, yielding an output vector *Wi*. When the multiple runs are completed, the weight *multi_W*(*s*) of a position *s* is computed by averaging the positive weights assigned to that position by the *nr_iter* runs of RELIEF. That is, using $N^+=|\{i|W_i(s)>0\}|$ and $N^-=|\{i|W_i(s)<0\}|$

$$multi_W(s)=\begin{cases}\dfrac{1}{N^+}\sum_{\substack{i\\W_i(s)>0}}W_i(s) & \text{for} \quad N^+>0\\[2ex] \dfrac{1}{N^-}\sum_{\substack{i\\W_i(s)<0}}W_i(s) & \text{for} \quad N^+=0\wedge N^->0\\[2ex] 0 & \text{for} \quad N^+=0\wedge N^-=0\end{cases}$$

Note that in the definition of *multi_W*(*s*), only those runs where RELIEF assigned a positive weight to *s* are considered. In this way, *multi_W*(*s*) assigns a high score to position *s* only if it discriminates at least two classes. In particular, a maximum weight will be assigned if *s* fully discriminates two specific classes but does not differentiate (i.e., weight less than or equal to zero) any other pair of classes.

Random sampling of pairs of classes is mainly employed for efficiency reasons, while random sub-sampling of sequences is applied for handling unbalanced classes as well as for gaining efficiency. The computational complexity of multi-RELIEF is O(*nr_iter* * *nr_sample*^{2}* *nr_positions*) while that of RELIEF-F is O(*nr_seq*^{2}**nr_positions*). Algorithms that do not consider the context (univariate scoring algorithms), such as TEA and SH, are generally more efficient with complexity O(*nr_seq* * *nr_positions*).

Table 1 illustrates the application of multi-RELIEF to a toy dataset. Note that positions *b* and *c* both get maximum weight. This is expected for position *c*, because it fully discriminates

each pair of classes. Instead, position *b* only discriminates a subset of classes, e.g., *C1/C3*, while it does not separate other pairs of classes, i.e., *C1/C2* and *C3/C4*. So only residue positions that, at least partly, discriminate between pairs of classes have a positive weight assigned by multi-RELIEF. This property of the algorithm is desirable, e.g. in cases where the number of subfamilies is larger than the number of amino acids, such as the GPCR benchmark (see below) that consists of 77 classes.

Table 1. Weights computed by multi-RELIEF applied to a toy example

	a	b	c	d	e
C1	R	F	T	I	T
	R	F	T	Q	F
	R	F	T	N	V
	R	F	T	A	D
C2	R	F	Y	S	T
	R	F	Y	F	F
	R	F	Y	D	V
	R	F	Y	L	D
C3	R	Y	D	E	T
	R	Y	D	V	F
	R	Y	D	W	V
	R	Y	D	G	D
C4	R	Y	H	H	T
	R	Y	H	P	F
	R	Y	H	Y	V
	R	Y	H	C	D
Weights	0	1	1	0	-1

Multi-RELIEF+3D contacts

As an additional step in multi-RELIEF, 3D structural information can be exploited. We use a simple heuristic based upon the notion that functional specificity generally does not evolve for a single residue but typically involves a cluster of residues in the protein structure. For each position *s*, we adjust the corresponding multi-RELIEF weight by adding the average weight of its 3D neighbors. Thus, the score of a residue will be boosted if its neighbors have a high average score. 3D neighbors are residues that share surface with a given residue as calculated by the web server at http://ligin.weizmann.ac.il/cma/ (Sobolev et al., 1999). From the list returned, residue pairs with a sequence distance of two or less are removed.

Comparison to other algorithms

1. TEA: Ye et al. (2006) and also in Chapter 2
2. SDPpred: http://bioinf.fbb.msu.ru/SDPpred/index.jsp

3. TreeDet/MB: http://www.pdg.cnb.uam.es/Servers/treedet/
4. SH: http://www/ibi.un.nl/programs/seqharmwww/

To be able to use TEA automatically we used the following scoring function:

$$score(s) = Entropy(s, D) - \frac{1}{N} \sum_{C \in Classes} Entropy(s, C)$$

for each position *s* in a given MSA *D* partitioned into *N Classes C*, and using *Entropy(s, X)* for the entropy of *s* computed on dataset *X* (See Chapter 2). SDPpred was applied with 10000 shuffles for each column, and a maximum allowed percentage (70%) of gaps in a group in each column; these are the highest possible settings allowed through the web interface of SDPpred. TreeDet/MB was applied with the following setting, in order to obtain a ranking of the residues: advanced run for MB method, cutoff set to 10^{-12} and percentage of High Scoring Residues set to 100%. We could not run TreeDet on the GPCR dataset because its web server accepts a maximum of 200 sequences. For this reason, we compiled a GPCR-190 reduced set (see below), to which TreeDet was applied. *SH* has no adjustable parameters, except for the cutoff value that is irrelevant for the generation of the ROC curves used. Note that for a fair comparison between the methods, the tie-braking by sequential groups ("Rank") and entropy was excluded from the *SH* method. A similar mechanism could be added to the other methods in a postprocessing step. *SH* was not applied to the GPCR and LacI datasets since these consist of more than two classes.

Multi-RELIEF was run using parameters *nr_iter*=1000 and *nr_samples*=10. These values were chosen based on the number of classes and their sizes, albeit no parameter tuning was applied. In general, a high value of *nr_iter* and a reasonably small value of *nr_samples* are recommended. Ties were broken by sorting residue positions with equal score in increasing sequence position.

Benchmark studies

The performance of a method may depend on the type of protein family and functional specificity properties considered. We therefore carried out a benchmark involving seven different protein families with various associated functional specificity properties (Table 2).

G Protein-Coupled Receptors (GPCRs) are integral cell membrane proteins involved in signal transduction. Their mediatory role makes them important drug targets (Gether et al., 2002; Pierce et al., 2002). We extracted the MSA of class A GPCRs in the transmembrane region from the latest version of the GPCRDB (Horn et al., 2003, June 2006 release 10.0, http://www.gpcr.org/7tm/), yielding an MSA of 2065 protein sequences with an average identity of 26% over all sequence pairs in the alignment. The MSA was classified into 77 subfamilies according to the recognition of endogenous ligands. An additional reduced MSA was derived by applying a redundancy limit of 65% identity, and subsequently discarding all

subfamilies that had only one sequence remaining. This yielded an MSA of 190 protein sequences divided over 39 GPCR families, which was named “GPCR-190”. Residues are deemed to be ligand binding whenever their mutation affects ligand binding in aminergic receptors, as listed in Table 2.

The LacI family is one of the largest families of bacterial transcription factors. This family was analyzed by Mirny and Gelfand (2002) using a technique based on mutual information. We used a multiple sequence alignment of 54 LacI protein sequences (Mirny and Gelfand, 2002) classified into 15 families. Suckow et al. (1996) mutated positions 2–329 of Lac repressor into 12 or 13 of the 20 natural occurring amino acids. These 4000 well-defined mutants yielded a functional classification for each position. We took the residues in group IX (DNA binding) and XI (IPTG binding) as the specificity residues. Some of these are actually conserved in the alignment and thus cannot contribute to specificity. These were subsequently excluded from the selection. The resulting 28 specificity determining residues are listed in Table 2.

The Ras superfamily of small GTPases is implicated in the regulation of growth, survival, differentiation and other processes in haematopoietic cells (Reuther and Der, 2000). It comprises six families, of which experimental evidence for functional sites was available from the literature for the Rab 5 versus Rab 6 subfamilies, and the Ras versus Ral families, as defined in Pirovano et al. (2006). The 28 and 12 true positives, respectively, are listed in Table 2. The MSAs of 4 Rab5 and 6 Rab6, and of 20 Ras and 69 Ral protein sequences described in Pirovano et al. (2006) were used.

The Major Intrinsic Protein (MIP) family of Integral Membrane Transporters is mainly involved in facilitating the transport of both water and small neutral solutes through the cellular membrane in all domains of life. There are about six MIP subfamilies, the two major are the aquaporins (AQPs) and the glycerol-uptake facilitators (GLPs) (Zardoya and Villalba, 2001). The MSA of 12 AQP and 48 GLP protein sequences described in Pirovano et al. (2006) was used. Residues with at least one atom closer than 7 Å to the bound glycerol molecules in the GLP pore channel in the crystal structure 1FX8 (Fu et al., 2000), excluding those that were conserved in the training set of sequences, as defined in Pirovano et al. (2006). This yields a set of 37 sites, which are listed in Table 2.

The Smad family of TGF β-associated transcription factors plays a crucial role in the transforming growth factor-β signaling pathway and is critical for determining the specificity between alternative pathways (Feng and Derynck, 2005; Massague et al., 2005). The family can be subdivided into two major classes: AR-Smads that are mainly induced by TGF_-type receptors, and BR-Smads that are mainly induced by the BMP-type receptors. The MSA of 8 AR-Smad and 12 BR-Smad nonredundant sequences of the Smad-MH2 domain described in Pirovano et al. (2006) was used. The 29 specificity determining residues as defined in Pirovano et al. (2006) are listed in Table 2.

Table 2. Properties of the datasets used for testing the algorithms

Dataset	Number of classes	Average (SD) class size	Max, min class size	Number of positions in MSA	Site information	"True" sites
GPCR	77	26.8(34)	189, 3	214	Ligand	T94, T97, E113, G114, A117, T118, G121, L125, C167, L172, F203, V204, M207, F208, H211, Y268, A269, A272, A292, F293
GPCR-190	39	4.9(3.9)	21, 2			
LacI	15	3.6(2.5)	12, 2	339	Ligand and DNA	T5, L6, S16, Y17, Q18, R22, N25, Q26, H29, Q54, A57, S61, L73, A75, P76, I79, N125, P127, D149, S191, S193, W220, N246, Q248, Y273, D274, T276, F293
Ras Ral	2	44.5(24.5)	69, 20	218	Protein	I24, Q25, D30, E31, D33, I36, E37, Q43, L53, M67, Q70, D92
Rab5 Rab6	2	5.0(1)	4, 6	163	Protein	K21, G22, Q23, H25, E26, F27, Q28, E29, S30, H62, A65, M67, Y69, G71, A72, Q73, E96, L97, Q98, R99, Q199, A101, S102, P103, N104, I105, V106, K162
AQP GLP	2	30.0(18)	48, 12	430	Protein	L21, W48, V52, A65, H66, L67, V71, T137, T138, P139, N140, P141, L159, I163, I187, G195, P196, L197, G199, F200, A201, M202
Smad	2	10.0(2)	12, 8	211	Protein	L263, Q264, T267, Q284, Q294, P295, L297, T298, S308, E309, A323, V325, M327, I341, F346, P360, Q364, R365, Y366, W368, N381, R427, T430, S460, V461, R462, C463, M466

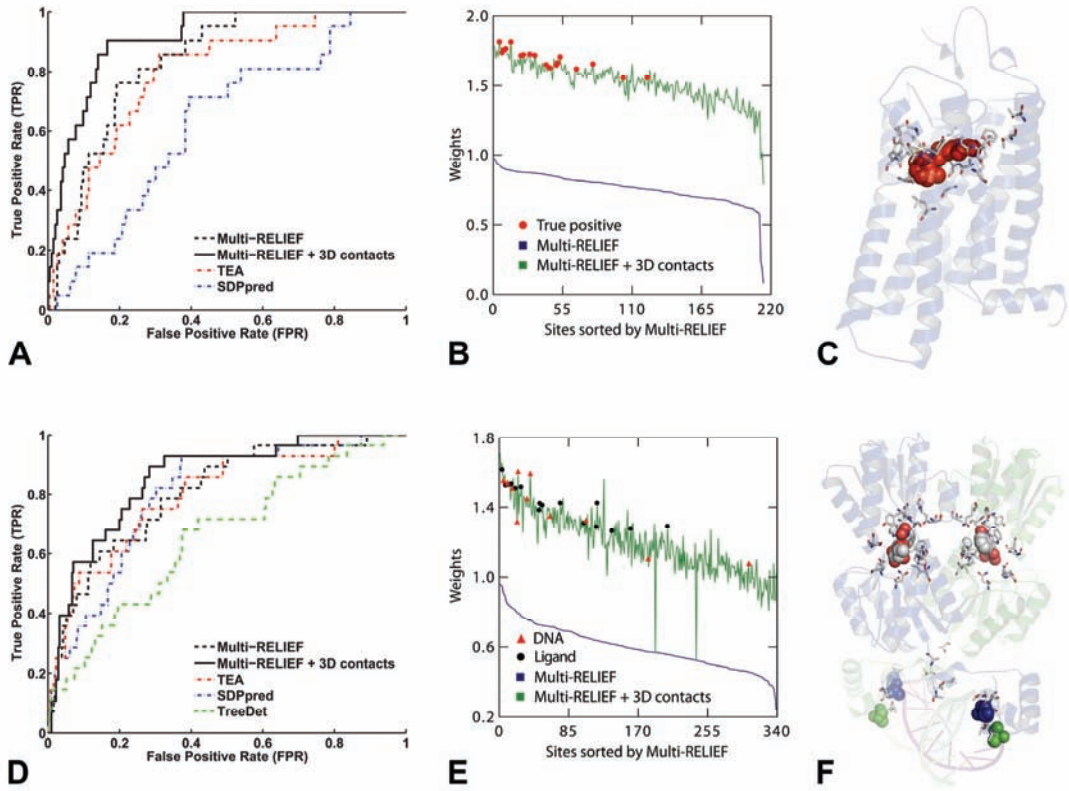


Figure 1. Results for all methods on the GPCR (A–C) and LacI (D–F) datasets. In columns are the ROC curves (A and D); the weights assigned by multi-RELIEF without (blue) and with (green) 3D contacts, and true positives highlighted with symbols (B and E); and the respective protein structures with true positive residues in sticks, and ligands in space filling balls (red for GPCR, C, and atom colors for LacI, F). Note that TreeDet could not be applied to the GPCR dataset (A) due to its size (4200 sequences). For LacI, the residues S21 and A27 mentioned in the text are highlighted as blue spheres (F).

Evaluation of the algorithms' performance

The Receiver-operator characteristic (ROC) curve is used for testing the performance of an algorithm for separating true and false positives (Provost and Kohavi, 1998; Swets, 1988). Here known functional specificity residues are considered true positives. The remaining residues are considered true negatives. We use the scoring (weight) values as threshold for generating the ROC curve. For each weight value *v*, the set of residues with weight higher than or equal to *v* is considered: the true positive percentage is reported on the *y*-axis (sensitivity, or coverage), and the false positive percentage (1-specificity, or error) on the *x*-axis. The ROC curve thus describes the goodness of a method in giving higher ranking to the *given* functionally important residues.

Results

From the ROC curves in Figure 1A and D, the results of our multi-RELIEF method appear superior to the other methods over all the datasets. The addition of 3D contacts yields a clear

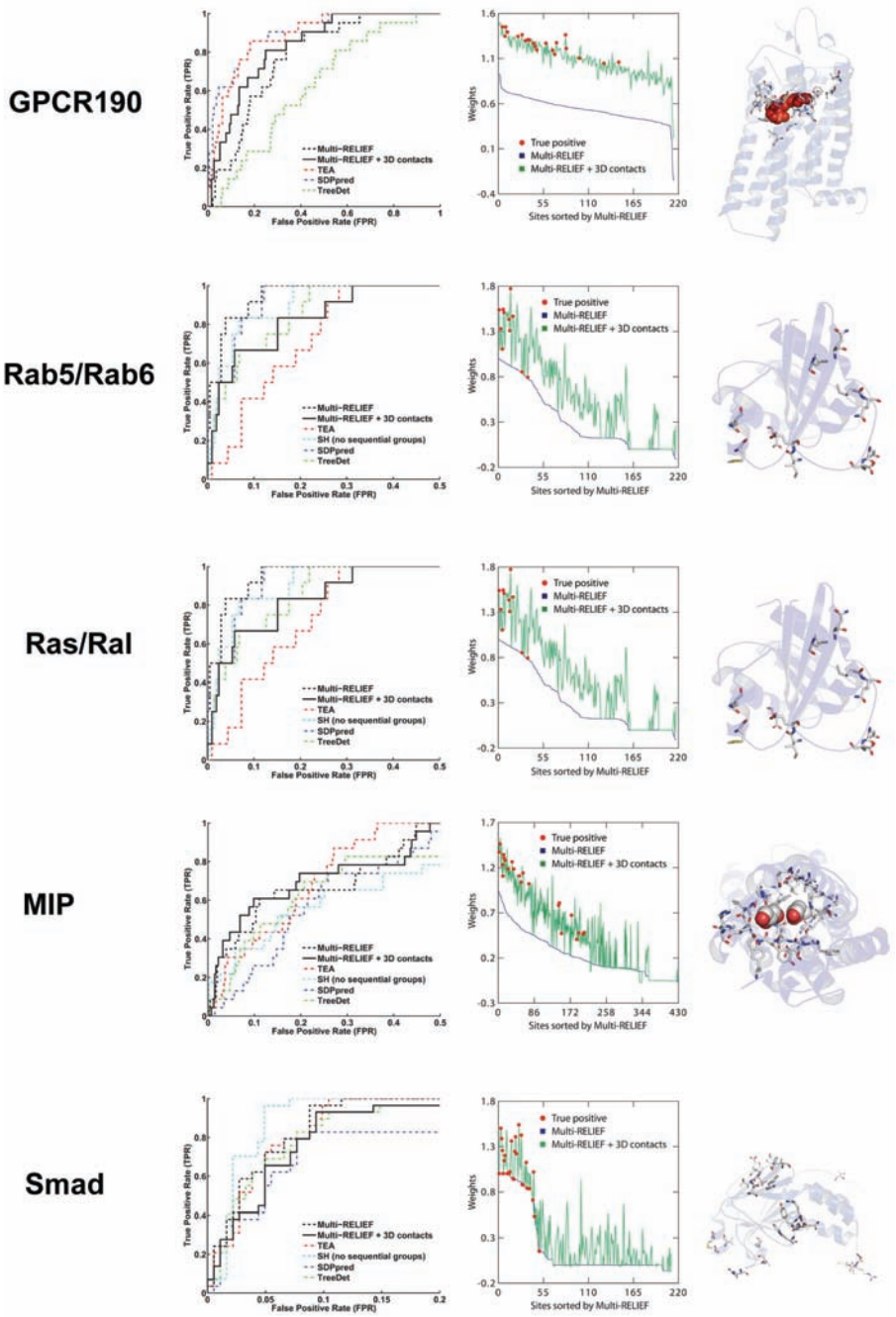


Figure 2. Results for additional datasets and all methods, cf. Figure 1. In columns from left to right are: ROC curves (note the differences in scale of FPR axis); The weights assigned by multi-RELIEF without (blue) and with (green) 3D contacts, true positive residues for ligand binding are highlighted in the latter curve (red circles); and the respective proteins with true positive residues in sticks, and ligand (in any) space filling balls.

improvement for the GPCRs and LacI family, as is shown in the “weights” plots in figure 1B and E. This is more evident for the GPCRs, for reasons explained in the next section.

The Smad, Ras/Ral and Rab5/Rab6 datasets contain two classes, which are rather balanced. In this case nearly all algorithms achieve high performance, but some variations are still observable (see Figure 2). In general, the distributions of true positives with respect to the computed scoring weights are similar for Smad, GPCR and Ras/Ral, and for Rab5/Rab6 to somewhat lesser extent. The true positives occur in the upper part of the curve, i.e., they satisfy the multi-RELIEF condition of being locally specific.

For the LacI and MIP datasets the situation is slightly different. Here, the majority of true positives also occur on the upper part of the curve, but some are retained at the central or lower parts of the curve. Clearly, some of the LacI true positives do not confirm to the model of local specificity exploited by multi-RELIEF. Upon detailed examination, these sites turn out to be largely conserved, *a*-like positions, as discussed further below.

The overall performance of the methods can be captured by the area under the curve (AUC) in the ROC plots, as listed in Table 3. Here we observe that in five of the datasets, multi-RELIEF or multi-RELIEF + 3D contacts is the best-scoring algorithm. In two others, they are not far below the best. A notable exception is the GPCR-190 reduced set, for reasons that are explained below. Importantly, the other methods are top-scoring only in at most one single dataset. The average scores over all datasets, in the last column of Table 3, also shows that multi-RELIEF and multi-RELIEF + 3D contacts are the top-scoring methods, with a consistent but modest lead for multi-RELIEF + 3D contacts.

Discussion

Evolving Specificity Residues

It is well accepted in sequence analysis that conserved residues are likely to be functionally important. Indeed, many early approaches select functional sites by simply picking the most conserved positions in a given MSA. Since vast amounts of sequence data have become available, sequence comparison between paralogous and orthologous proteins is performed routinely in order to identify specificity residues that account for differences between functional subgroups. Most state-of-the-art approaches for functional specificity detection require an MSA with predefined functional classes. They then forward MSA positions conserved within each group but different between groups as functionally specific. However, different degrees of specificity may be relevant. For example, position *c* in the Table 1 (the toy example) provides a perfect explanation of such subdivision in classes. Although position *b* is insufficient for differentiating all four classes, it does provide some information about the difference between *C1*, *C2* and *C3*, *C4*. We refer to these type of positions as a-positions, b-positions, and so on.

The specificity residues considered in this chapter include *c*-like positions, that are fully class-specific, but also partially class specific *b*-like positions are present, especially in the GPCR

case study. The following evolutionary scenario can explain this observation. After proteins "learn" how to fold correctly in order to perform their main function, they can start evolving new functional sites in order to interact with other components such as small molecules, DNA, RNA or another protein. Such a process can be conducted in a stepwise fashion, first by establishing general interaction anchor points (conserved, *a*-like positions), next by evolving to more selective recognition sites (specific, *b*- and *c*-like). For example, if proteins want to interact with DNA, they first evolve some positively charged residues in a certain region of the protein just to attract the negatively charged phosphoric acid group(s) of DNA. They then can evolve *b*-like positions to selectively bind to a specific category of DNA and finally, they can obtain *c*-like positions to achieve specific recognition of a particular DNA fragment.

Table 3. Area under curve for the ROC plots of the six methods and seven datasets, and average scores relative to multi-RELIEF over the common datasets (best scores in bold)

Method	GPCR	GPCR-190	LacI	Rab5/Rab6	Ras/Ral	AQP/GLP	Smad	Relative average
Multi-RELIEF	0.83	0.78	0.80	**0.90**	**0.97**	0.83	0.97	0
+3D	**0.91**	0.84	**0.85**	0.86	0.91	**0.84**	0.96	**+0.003**
TEA	0.80	0.89	0.80	0.79	0.86	**0.84**	0.96	-0.039
SH	-	-	-	0.86	0.95	0.75	**0.98**	-0.033
SDPpred	0.63	**0.90**	0.80	0.83	0.96	0.78	0.84	-0.058
TreeDet/MB	-	0.63	0.66	0.85	0.92	0.79	0.96	-0.073
Average	0.79	0.81	0.78	0.85	0.93	0.81	0.95	

Functional specificity sites can therefore contain different types of specificity positions. The proportions of *a*-, *b*- and *c*-like positions (see Table 1) may vary within different protein families. In our benchmark studies, for the GPCRs, Smad and LacI datasets, we defined all residues at the specificity interaction interface according to the experimental evidence and excluded a-like. Such definition is straightforward but results in *c*- and *b*-like positions being taken as "true" positives. If a family contains a high percentage of *c*-like positions, methods focusing on intra-group conservation will all perform well, while a more varying performance is likely with larger proportions of *b*-like positions.

Using 3D contacts

Although multi-RELIEF attains similar or better performance than its counterparts considered here, we have demonstrated that specificity detection can be further enhanced by taking 3D contact information into account (Figure 1A and D). In this scenario, the score of a residue position will be boosted if its neighboring residues score high, introducing a bias towards spatially clustered residues. Depending on the ligand being a small molecule or a larger

protein or DNA structure, employing contact information may affect predictions differently. If the ligand is a small moiety, the specificity residues form a small, compact cavity, such that application of 3D contacts improves prediction. On the other hand, interaction interfaces to large protein or DNA ligands will generally be larger and more planar, often leading to relatively few isolated interface residues providing specificity recognition. This renders 3D contacts potentially less beneficial for datasets associated with proteins interacting with larger ligands.

Benchmark performance

GPCRs On the GPCRs dataset, all algorithms except SDPpred perform well. The GPCR ligand binding site is illustrated by retinal, the endogenous ligand of bovine rhodopsin in Figure 1C. Multi-Relief outperforms the other methods substantially, and the use of structure information (multi-Relief+3D contacts) further improves its performance. There are two factors that contribute to these observations. First, there are 77 subfamilies in the GPCRs dataset that cannot be uniquely differentiated by a single position using the 20 natural amino acids. Thus, in the absence of absolute *c*-like positions, *b*-like positions are the best alternative. This gives multi-Relief an obvious advantage in identifying b-like positions. Second, the class A GPCRs evolved to recognize small molecules so that the specificity site is relatively compact and concentrated in a small region of the protein compared to other protein families that recognize DNA, RNA or protein (Figure 1C). This also explains the relatively large performance increase, compared to the other datasets, of multi-Relief when 3D contacts are used for boosting results for the GPCR dataset.

For the GPCR-190 set, average AUC of all methods is similar to that of the full GPCR set, see Table 3. Intriguingly, only multi-RELIEF performs similarly over both datasets, while all other methods perform differently. Multi-RELIEF+3D contacts give a much smaller improvement over multi-RELIEF than in the full GPCR set, but more strikingly, the performance of TEA and SDPpred are higher. An explanation can be found in the 65% redundancy threshold applied. This retains diversity within a subfamily, i.e., the most divergent members, but multi-RELIEF relies on differences between nearest neighbors, which could be entirely different in the reduced set. Even the 3D information apparently cannot overcome this. TEA and SDPpred, on the other hand, put more emphasis on entropy to measure the overall differences between the subfamily, which may be more pronounced in the reduced set.

LacI Results on the LacI dataset highlight the difference between specificity-related binding to small molecules compared to binding DNA. LacI transcription factors bind to particular DNA fragments to prevent transcription of downstream genes. After recognition of ligands specific for each subfamily, they change conformation so that RNA polymerase is no longer blocked from binding to DNA. This leads to high expression of the encoded proteins. As illustrated in Figure 1E, multi-RELIEF generally assigns higher weights to residues that

recognize the small molecule than those binding to DNA. Moreover, application of the 3D contacts option boosts the weights of these residues.

Figure 1F shows the structure of the transcription factor of LacI. Among the small molecule binding sites, the position of R196 has low weight, even after being boosted by means of the 3D contacts step. When looking at its residue composition (data not shown), we can see that R196 is a *b*-like position, since amino acid R occurs in 47 out of a total of 54 protein sequences.

For this dataset, the 3D contacts information does not notably improve the detection of the DNA-binding residues, but, importantly, the prediction quality also does not suffer from the 3D contacts. The limited added value may be due to the fact that the DNA binding site is much bigger and more extended than the binding site for small molecules. Thus, interaction between protein and DNA may include several relatively isolated and spatially separated locations. For example, residue S16 interacts with DNA and indeed is assigned a high weight since it contributes to specific recognition. However, neighboring residues do not interact with the DNA and have low weight, so the score of S16 becomes worse after application of the 3D contacts step.

In addition, we identified a specific region of the protein where two residues, S21 and A27 are close to each other and have high weights before and after application of 3D contacts. Although these two residues were not characterized as DNA binding by Suckow et al. (1996), they are located within 5 Å distance to the DNA.

Ras The two datasets from the Ras family are based upon mutation experiments, three regions of about 10 residues each for Rab5 versus Rab6, and 12 point mutations for Ras versus Ral and Rab. They show best performance for multi-RELIEF and worst for TEA, while other methods perform very similar and only slightly below multi-RELIEF, see Table 3 and Figure 2. Overall performance of all methods is lower for Rab5/Rab6 than for Ras/Ral (Table 3).

Although specificity in the Ras superfamily is related to recognition of various small-molecule and protein targets, multi-RELIEF is well able to recognize these sites. However, due to the presence of multiple interacting sites, addition of 3D contacts information does not lead to a gain in detection of specificity residues.

MIP The MIP dataset is based on a structural definition of functional residues: those close to the ligand in the crystal structure. Overall performance of all methods is relatively low (see Table 3). Multi-RELIEF+3D contacts and TEA together are the best-scoring methods. Importantly, multi-RELIEF and multi-RELIEF+3D contacts show the steepest initial slope in the ROC curves, which is relevant for experimental planning if only top-scoring sites are to be examined.

Smad The Smad dataset is a special benchmark because the true positive residues have been verified directly by site-directed mutagenesis experiments. It is different from the other

datasets in that it contains two classes. The known functional specificity sites are a mix of *b/c*-like positions, i.e., specific and conserved in each class, and *d*-like positions, that are specific but not conserved within the classes.

The performance of all methods on the Smad is remarkably good, compared to the other datasets, see Table 3 and Figure 2 (note the difference in scale of the FPR axis). This is likely due to the comprehensive experimental coverage of true functional Smad sites, reducing the proportion of false negatives and increasing overall performance by all methods. The 3D contact step results in a slightly decreased performance of multi-Relief. This may be due to the fact that three different functional interactions are involved, each involving distinct interaction interfaces on the Smad protein surface.

CONCLUSION

In this chapter, we proposed a novel multi-RELIEF algorithm for identifying specificity-related functional sites. We provided an option for boosting prediction quality using structural information, if available, for specificity of interaction with small molecules. We tested the performance of multi-RELIEF and other recent algorithms on seven different experimental benchmark cases. The results demonstrate robustness and best overall performance of multi-RELIEF over a wide variety of biological cases.

REFERENCES

Bickel,P. et al. (2002) Finding important sites in protein sequences. Proc. Natl Acad. Sci. USA, 99, 14764-14771.

Carro,A. et al. (2006) Treedet: A web server to explore sequence space. Nucleic Acids Res, 35, 99.

DelSol Mesa,A. et al. (2003) Automatic methods for predicting functionally important residues. J. Mol. Biol., 326, 1289-1302.

Feenstra,K. et al. (2007) Sequence harmony: detecting functional specificity from alignments. Nucleic Acids Res., 35, W495-W498.

Feng,X. and Derynck,R. (2005) Specificity and versatility in TGF-beta signaling through Smads. Annu. Rev. Cell Dev. Biol., 21, 659-693.

Fu,D. et al. (2000) Structure of a glycerol-conducting channel and the basis for its selectivity. Science, 290, 481-486.

Gether,U. et al. (2002) Structural basis for activation of G-protein-coupled receptors. Pharmacol. Toxicol., 91, 304-312.

Gu,X. (2006) A simple statistical method for estimating type-ii (cluster-specific) functional divergence of protein sequence. Mol. Biol. Evol., 23, 1937-1945.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. J. Mach. Learn. Res., 3, 1157-1182.

Hannenhalli,S. and Russell,R. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. J. Mol. Biol., 303, 61-76.

Horn,F. et al. (2003) GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res., 31, 294-297.

Kalinina,O. et al. (2004) SDPpred: A tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. Nucleic Acids Res, 32, W424-W428.

Kononenko,I. (1994) Estimating attributes: Analysis and extensions of Relief. European Conference on Machine Learning, volume LNCS 784. Springer-Verlag New York, Secaucus, NJ, USA, pp. 171-182. http://portal.acm.org/citation.cfm?id=188427.

Landgraf,R. et al. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J. Mol. Biol., 307, 1487-1502.

Marchiori,E. et al. (2006) A feature selection algorithm for detecting subtype specific functional sites from protein sequences for Smad receptor binding. In Proceedings of the Fifth International Conference on Machine Learning and Applications (ICMLA'06), pp. 168-173. http://doi.ieeecomputersociety.org/10.1109/ICMLA.2006.7.

Massague,J. et al. (2005) Smad transcription factors. Genes Dev., 19, 2783-2810.

Mihalek,I. et al. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. J. Mol. Biol., 336, 1265-1282.

Mirny,L. and Gelfand,M. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J. Mol. Biol., 321, 7-20.

Pierce,K. et al. (2002) Seven-transmembrane receptors. Nat. Rev. Mol. Cell Biol., 3, 639-650.

Pirovano,W. et al. (2006) Sequence comparison by sequence harmony identifies subtype specific functional sites. Nucleic Acids Res.,34, 6540-6548.

Provost,F. and Kohavi,R. (1998) Guest editors' introduction: On applied research in machine learning. Mach. Learn., 30, 127-132.

Reuther,G. and Der,C. (2000) The Ras branch of small GTPases: Ras family members don't fall far from the tree. Curr. Opin. Cell Biol., 12, 157-165.

Robnik-Sikonja,M. and Kononenko,I. (2003) Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn., 53, 23-69.

Shenkin,P. et al. (1991) Information-theoretical entropy as a measure of sequence variability. Proteins, 11, 297-313.

Sobolev,V. et al. (1999) Automated analysis of interatomic contacts in proteins.Bioinformatics, 15, 327-332.

Suckow,J. et al. (1996) Genetic studies of the lac repressor. xv: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J. Mol. Biol., 261, 509-523.

Sun,Y. and Li,J. (2006) Iterative relief for feature weighting. In Proceedings of the 23rd International Conference on Machine Learning. ACM Press, USA, pp. 913-920.

Swets,J. (1988) Measuring the accuracy of diagnostic systems. Science, 240, 1285-1293.

Whisstock,J. and Lesk,A. (2003) Prediction of protein function from protein sequence and structure. Q. Rev. Biophys., 36, 307-340.

Ye,K. et al. (2006) A two-entropies analysis to identify functional positions in the transmembrane region of class a G protein-coupled receptors. Proteins, 63, 1018-1030.

Zardoya,R. and Villalba,S. (2001) A phylogenetic framework for the aquaporin family in eukaryotes. J. Mol. Evol., 52, 391-404.

Chapter 5

An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences

Motivation: Pattern discovery in protein sequences is often based on multiple sequence alignments (MSA). The procedure can be computationally intensive and often requires manual adjustment, which may be particularly difficult for a set of deviating sequences. In contrast, two algorithms, PRATT2 (http//www.ebi.ac.uk/pratt/) and TEIRESIAS (http://cbcsrv.watson.ibm.com/) are used to directly identify frequent patterns from unaligned biological sequences without an attempt to align them. Here we propose a new algorithm with more efficiency and more functionality than both PRATT2 and TEIRESIAS, and discuss some of its applications to G protein-coupled receptors, a protein family of important drug targets.

Results: In this chapter, we design and implement six algorithms to mine three different pattern types from either one or two datasets using a pattern growth approach. We compare our approach to PRATT2 and TEIRESIAS in efficiency, completeness and the diversity of pattern types. Compared to PRATT2, our approach is faster, capable of processing large datasets and able to identify so-called type III patterns. Our approach is comparable to TEIRESIAS in the discovery of so-called type I patterns but has additional functionality such as mining so-called type II and type III patterns and finding discriminating patterns between two datasets.

Based upon Kai Ye,Walter A. Kosters and Adriaan P. IJzerman
Bioinformatics 23: 687-693 (2007)
The source code is available at http://www.liacs.nl/home/kosters/pg/

Introduction

One of the crucial topics in the analysis of biological data is the discovery of frequent patterns in a set of DNA or protein sequences. These patterns usually hint at shared biological functions. For example, some patterns may be essential for the proteins to fold correctly while other patterns may form a certain micro-environment to recognize a small molecule ligand or another protein partner. Various algorithms have been designed to identify such patterns either by overlaying protein structures (Copley et al., 2001; Lupas et al., 2001; Russell et al., 1998) or by mining in sequences. For the latter, most pattern discovery approaches either use aligned sequences as an input or create a multiple sequence alignment (MSA) in an early stage of the analysis such as PRINTS (Attwood et al., 2003), PROSITE (Hulo et al., 2006) and Pfam (Baldi and Chauvin, 1994; Baldi et al., 1994; Bateman et al., 1994; Shigeta et al., 2003). In addition to MSA, some algorithms such as correlation-based approaches (Kuipers et al., 1997), evolutionary trace (Lichtarge et al., 1996) and two-entropies analysis (Chapter 2) even include a phylogeny to uncover further information about functional sites. Construction of such multiple sequence alignments, however, requires parameterization, is computationally intensive, often requires manual adjustment, and can be particularly difficult for a set of deviating sequences.

In contrast, TEIRESIAS (Rigoutsos and Floratos, 1998) and PRATT2 (Jonassen, 1995; Jonassen et al., 1997) are used to directly identify frequent patterns from biological sequences without aligning them. The combinatorial pattern discovery algorithm TEIRESIAS identifies patterns in two stages. In the first stage of scanning the elementary patterns, TEIRESIAS identifies all elementary patterns of length at most W, with at least L non-wildcard items ($W \geq L$). The elementary patterns must show up in at least K sequences of the input, the so-called support of the pattern. In the second stage of elementary pattern convolution, all elementary patterns are combined into larger patterns until no more patterns emerge. Although TEIRESIAS reports all patterns which fulfill the predefined parameters and does not score and rank patterns, it provides utilities in its web server to filter out some patterns based on properties such as the number of non-wildcard items in the patterns. Using a pattern graph, PRATT2 searches for conserved patterns showing up in at least k out of n sequences. It uses the $(n-k+1)$ shortest sequences to construct a pattern graph and then uses the pattern graph to mine patterns. The identified patterns are scored and only the top 50 patterns are reported by default. PRATT2 can also identify patterns with limited flexibility in the number of consecutive wildcards. In the additional refinement stage, it replaces some wildcards with ambiguous residues if it can find any.

In this chapter, we follow the ideas of pattern growth that are emerging in computer science (Pei et al., 2004) and design six algorithms to provide a complete solution of pattern discovery in unaligned protein sequences. In computer science, quite a few studies have contributed to the efficient mining of sequential patterns. Almost all of them are Apriori-like,

i.e., based both on the Apriori principle, which states that any super-pattern of an infrequent pattern cannot be frequent, and on a candidate-generation-and-test approach (Agrawal et al., 1994). One of the most efficient algorithms is PrefixSpan (Pei et al., 2004), which mines sequences of itemsets. PrefixSpan has been used to mine very diverse datasets such as customer purchase patterns, web access patterns or disease treatments. However, PrefixSpan cannot be directly used to mine frequent patterns in biological databases because it was designed to mine sequences of itemsets instead of sequences of items (DNA or protein sequence) and report computer style patterns without any constraints on gap length. For example, a typical input sequence for PrefixSpan may be (acd)(edh)(aij)(ahi) in which (acd) is an itemset. PrefixSpan may report a pattern like a*d*i in which "*" means an undefined number of wildcards. Regular expression constraints were first introduced in the Apriori framework to find frequent patterns with user-predefined items (Garofalakis et al., 2002). Later, Pei and coworkers introduced regular expressions into the mining process of PrefixSpan as well as six other constraints such as a timestamp difference between every two adjacent frequent items (Pei et al., 2002). However, in a biologically meaningful pattern, not only frequent items and their sequential order but also the number of wildcards between frequent items carries essential information. None of the present extensions in sequential pattern mining allows finding the number of wildcards between frequent items so that they cannot yield biologically meaningful PROSITE-like patterns. In this chapter, we use the principles of pattern growth as present in PrefixSpan and introduce both wildcard constraints to mine PROSITE-like patterns and a sequence sliding window in the pattern growing process. We grew patterns continuously in a so-called projected database. The projected databases keep on shrinking by eliminating sequences which did not support the current growing pattern. The patterns stop growing if the current projected database is smaller than the predefined support threshold. In this way, we discover the complete pattern set efficiently.

From one dataset, we can identify three different types of patterns. Type I is a pattern with a defined number of wildcards such as AxxT, where x demotes a single wildcard. Type II is a PROSITE like pattern (Hulo et al., 2006). For example, Ax(3,4)DF is a pattern with three residues and one limited flexible wildcard region. It matches any protein sequence with "A" followed by three or four residues or wildcards and ending with DF. Type III patterns match protein sequences which have the residues placed in the right order within a predefined length (window). For example, a type III pattern A*T*D with a window of 8 matches protein sequences with A, T and D placed in that order in an 8 residues peptide fragment. The concept of length constraint (sliding window) for sequential pattern mining was first proposed by Pei et al. (2002).

We compared our program with PRATT2 and TEIRESIAS in identifying the type I patterns since all three are able to find this type of patterns. After that we compared our program with PRATT2 in finding type II patterns. We also investigated the performance of our program in

mining the type III patterns which are novel in protein sequence analysis. In addition, we adapted our algorithms to mine two datasets at the same time. This allowed us to identify one of the above pattern types to discriminate between the two datasets.

Methods

Algorithms

A *pattern* is an ordered series of *items*, where each item is either one of 20 residues or 3 special *wildcard* characters, x, x(*i*,*j*) or * in which x matches any of 20 residues, x(*i*,*j*) matches *i* to *j* ($0 \leq i \leq j$) residues, * matches any combination of residues with undefined length. The *support* of a pattern in a database is the ratio of sequences that satisfy the pattern over all sequences. Note that if a pattern occurs twice or more in a sequence, it is only counted once.

Since the algorithms for mining the three types of patterns from one dataset are very similar, we describe the algorithm for type I patterns from one dataset (PGOneI, PG stands for pattern growth) in detail. The descriptions of the other two, PGOneII and PGOneIII are available at http://www.liacs.nl/home/kosters/pg/. A method similar to PGOneIII, in which a length constraint is introduced into the sequential pattern mining process, was also described by Pei et al. (2002).

The inputs of the algorithm are dataset *S* which consists of a series of non-empty sequences, a minimum support threshold *min_support*, the minimal number of non-wildcard items in the patterns to be reported *min_non_wc*, and the maximal number of consecutive wildcards x *max_wc_l*.

The output of the algorithm consists of all patterns that have support larger than or equal to *min_support*, and that satisfy the other input parameters. The supports are also given. These patterns are called *frequent*.

The algorithm is as follows, see the next page. Here *a* is a pattern and S_a is the so-called projected database that contains all sequences that satisfy the pattern *a*, where the last element of each occurrence of *a* is marked (the so-called *a*-locations). Note that $|S_a|$, the number of sequences in S_a, is the support of *a*. Let *non_wc* (*a*) be the number of non_wildcard items in *a*, *curr_wc_l* (*a*) the number of consecutive wildcards at its end, and *last* (*a*) its last item.

The computation of $S_{a'}$ from S_a requires, for each *a*-location, the check whether or not the residue on its right side (if any) equals the newly appended item (b or x). In case of equality this gives an *a'*-location in $S_{a'}$. Sequences without *a'*-locations are deleted.

The main call is *PGOneI* (Λ, S_Λ), where Λ is the empty pattern; note that each residue in database S_Λ is a Λ–location, including the position *before* each sequence. This call creates a projected database that marks all occurrences of a single residue b, reports all frequent patterns that begin with this b, and then proceeds to the next amino acid.

```
PGOneI (a, S_a)
    if |S_a| ≥ min_support then
        if non_wc (a) ≥ min_non_wc and last (a) ≠ x then
            report a, |S_a|
        for each residue b do
            a' := a with b appended to it
            PGOneI (a', S_a')
        if a ≠Λ and curr_wc_l (a) < max_wc_l then
            a' := a with x appended to it
            PGOneI (a', S_a')
```

We also adapted our algorithms to mine two datasets at the same time to identify one of the three pattern types to discriminate between the two datasets. Two sequence datasets (one defined as positive dataset and the other as negative) are required. These two datasets were recursively mined at the same time and the support difference was evaluated and compared with a predefined *min_support_diff*. Only when the number of non-wildcards is no less than the predefined minimal number of non-wildcards *min_wc_eva*, the evaluation of the support difference starts since the supports are high in both datasets in the beginning of the mining process. Only those patterns with support difference no less than *min_support_diff* between the positive dataset and negative dataset are reported.

Comparison with PRATT2 and TEIRESIAS

In life sciences, TEIRESIAS and PRATT2 are the two state-of-the-art algorithms that identify patterns from a set of unaligned protein sequences. Therefore we compared our PGOneI algorithm with these two programs with respect to the identification of type I patterns from one dataset. In addition, PGOneII was compared with PRATT2 (current version pratt2.1) in identifying type II patterns in one dataset.

The PRATT2 algorithm developed by Jonassen et al. was downloaded from http://iubio.bio.indiana.edu/soft/molbio/pattern/. The TEIRESIAS algorithm was obtained from http://cbcsrv.watson.ibm.com/. The implementations of our algorithms, PRATT2 and TEIRESIAS, were compiled/installed under Red Hat Enterprise Linux WS 3.0. All tests were performed on the same PC with Intel Pentium 4 CPU 3.20GHz, 1.00 GB of RAM and a Maxline III 7L250S0 250GB Hard Drive.

Datasets

G protein-coupled receptors (GPCRs) are important drug targets. They are membrane-bound, serpentine-like, proteins with 7 transmembrane (TM), alpha-helical, domains. A subdivision in three classes has been proposed of which class A (rhodopsin-like) is by far the largest. We used class A GPCRs as our test set for at least two reasons. First, class A GPCRs contain thousands of divergent protein sequences, so that we can study the properties of our

algorithms in various sizes of real sequences. Second, class A GPCRs includes two big divergent datasets (olfactory receptors and non-olfactory receptors) which facilitate demonstration of our algorithms in two datasets.

Protein sequences of class A GPCRs were extracted from the GPCRDB (March 2005 release (9.0); http://www.gpcr.org/7tm/). We removed orphan receptors and split the sequences into two datasets which happen to be of similar size: olfactory (2034) and non-olfactory (2027). This large number of proteins and two subsets with similar sizes facilitate benchmarking of our various implementations with different functionalities. It also helped us to compare our algorithms with both PRATT2 and TEIRESIAS which also aim to discover patterns from unaligned protein sequences. To illustrate the impact of dataset size on the performance of our algorithms and the above two programs, we randomly extracted protein sequences from the non-olfactory dataset to form a series of datasets with different sizes (20, 50, 100, 200, 300, 400, 500, 600, 1000, 2000).

Reference type I patterns

In order to evaluate pattern discovery by PRATT2, TEIRESIAS and PGOneI from one dataset, we defined reference type I patterns though MSA as a "golden standard". We extracted the MSA for the transmembrane domains of all non-olfactory proteins from the GPCRDB and calculated the Shannon entropy for each position. We considered the 32 positions with an entropy value less than 1.5 as the globally conserved positions based on our previous study in Chapter 2. Then we combined the most frequent residue at the defined conserved positions in each helix as a pattern.

The reference patterns (32 defined positions) are GxxGNxxV for helix 1, FLxxLAxADL for helix 2, SxxxLxxISxDRY for helix 3, FxxPxxxxxxY for helix 5, FxxCWxPF for helix 6, LxxxNSxxNPxxY for helix 7.

Results

Implementations of the pattern growth algorithm

Six programs were implemented to identify frequent patterns using the principle of pattern growth. Three of them, PGOneI, PGOneII and PGOneIII aim to mine frequent patterns from one dataset while the other three, PGTwoI, PGTwoII and PGTwoIII work at two datasets at the same time to find patterns that discriminate members of a positive dataset from those of a negative dataset. Figure 1 depicts an overview of our solutions to the problem of mining frequent patterns from unaligned protein sequences. "PG" stands for "**P**attern **G**rowth" since we use a pattern growth approach.

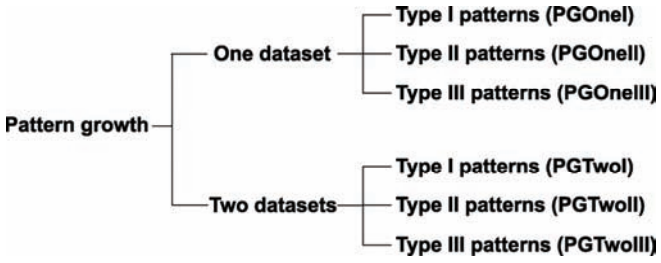


Figure 1. Overview of six implementations of the pattern growth approach.

Characteristics of PGOneI

Since PGOneI, TEIRESIAS and PRATT2 are all able to find type I patterns from one dataset, we compared these three algorithms in efficiency and completeness of pattern discovery.

We first examined the efficiency when we varied the size of input, the support and the maximum number of consecutive wildcards. For the size of input, a series of random data sets from the non-olfactory protein set with various sizes were used as inputs and the runtimes were recorded. The three programs were used to find type I patterns from one data set in two runs for different supports (0.3 for low support and 0.7 for high support). The maximal number of consecutive wildcards in the patterns was 3. The refinement step of PRATT2 was turned off to speed it up.

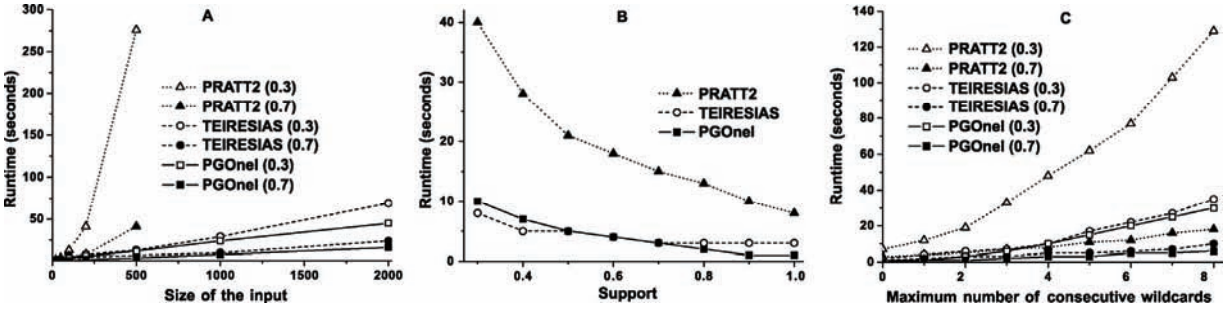


Figure 2. Characteristics of PGOneI and comparison of the runtime of PGOneI with PRATT2 and TEIRESIAS on different sizes of inputs, different supports and various maximum numbers of consecutive wildcards.

As shown in Figure 2A, both TEIRESIAS and PGOneI succeeded in all tests and the runtimes are almost linearly correlated to the size of the input. PGOneI is slightly more efficient than TEIRESIAS. The runtime of PRATT2 increases dramatically when the size of the input increases. PRATT2 fails when the input contains 600 proteins independent of the levels of the support. For all three programs, the runtime for searching patterns with high support is of course much shorter than for those with low support.

To test how support affects the efficiency of the three algorithms, we used a random set with 200 proteins of non-olfactory receptors as input. As shown in Figure 2B, the performance of PGOneI is comparable to TEIRESIAS and much better than PRATT2.

We also addressed how the number of consecutive wildcards affects the efficiency. We recorded the runtime of the three algorithms in search of patterns with a different maximum number of consecutive wildcards from a random set with 200 proteins of non-olfactory receptors. Figure 2C shows again that PGOneI is comparable to TEIRESIAS and much better than PRATT2.

Table 1. Completeness of pattern discovery for PRATT2, TEIRESIAS and PGOneI.

Program	Size of report	Number of positions identified out of 32
PRATT2	1,000	23
TEIRESIAS	2,742	28
PGOneI	927	28

To evaluate the completeness of PGOneI, PRATT2 and TEIRESIAS in finding type I patterns, we ran the three programs using the dataset with 500 non-olfactory proteins as input since PRATT2 failed at an input of 600 proteins. The three programs were configured to find type I patterns with the maximum number of consecutive wildcards equal to 3. The identified patterns were required to be present in at least 150 proteins. TEIRESIAS reported all patterns which satisfy the requirement. PGOneI yielded the same result as TEIRESIAS when we had it report all patterns (data not shown). However, since the probability of having a pattern with 2 non-wildcards in any protein of say 400 residues is close to 1, a lot of meaningless 2 non-wildcard patterns were diluting the pattern pool. As shown in Table 1, if we let PGOneI only report patterns with more than 2 non-wildcard residues, PGOneI yielded only about 1/3 of the patterns without missing any important patterns compared to TEIRESIAS. PRATT2 identified less functional positions because it ranks patterns by its scoring function and reports only the top ones (by default, PRATT2 reports the top 50 patterns and it is not able to report all patterns satisfying the parameters even when the maximal number of patterns to report is set big enough). Thus the mining results of PRATT2 are incomplete and important patterns may be missing. None of the three algorithms found all predefined 32 positions (see Section 2.4) for two reasons. First, the predefined maximum number of consecutive wildcards was equal to 3, prohibiting these three algorithms to find the pattern FxxPxxxxxxxY. Second, a combination of several most frequent residues may not be frequent enough to be detected, although the most frequent residue on an individual position is.

Characteristics of PGOneII

We implemented PGOneII to mine type II patterns from one dataset. Since TEIRESIAS cannot find type II patterns, we only compared PGOneII with PRATT2.

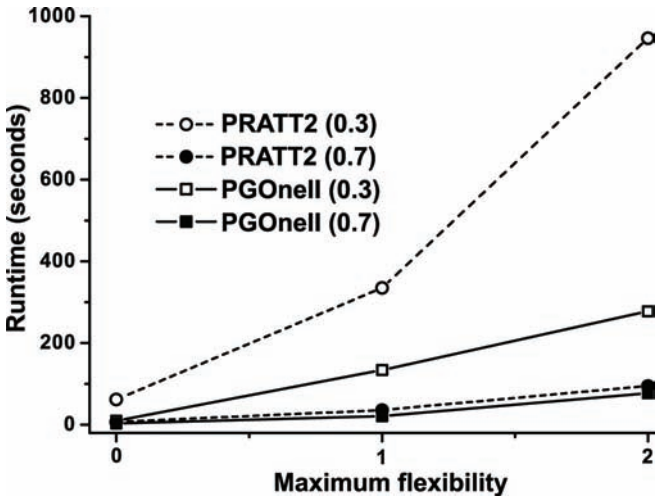


Figure 3. Characteristics of PGOneII and comparison with PRATT2 on maximum flexibility

We first used the dataset with 200 non-olfactory receptors as input to examine how the maximum flexibility affects the runtime of both PGOneII and PRATT2 in mining patterns with different support. The support values of 0.3 and 0.7 mean that the identified patterns must show up in at least 30% and 70% of proteins in the dataset, respectively. The maximum number of consecutive wildcards in the patterns was 3. The refinement step of PRATT2 was turned off to speed it up. We varied the maximum flexibility from 0 to 2. When the maximum flexibility is 0, no flexibility in the number of wildcards is allowed; if 1, patterns such as Ax(2,3)T and Wx(0,1)S are allowed. As shown in Figure 3, PGOneII is more efficient than PRATT2, especially in mining patterns with low support.

Then we compared PGOneII with PRATT2 on the different sizes of inputs, different supports and various maximum numbers of consecutive wildcards. The results were similar to the comparison between PGOneI and PRATT2 shown in Figure 2 (data not shown). For example, PRATT2 failed again before the size of the datasets reaches 600 proteins while PGOneII was successful in all runs.

We used the random set of 500 proteins (non-olfactory receptors) to compare the completeness of pattern discovery between PRATT2 and PGOneII. The maximum number of consecutive wildcards was set at 3, support 0.7 and the maximum flexibility 1. PGOneII identified 1,387 patterns. The maximum number of patterns to be reported by PRATT2 was set at 5,000 but PRATT2 reported only 1,111 patterns after it had scanned a total of 361,962 patterns. The patterns identified by PGOneII included all patterns found by PRATT2. PRATT2 failed to find some patterns such as Sx(2,3)Nx(0,1)PxxY which was identified by PGOneII and is well-known to be important for class A GPCRs.

Characteristics of PGOneIII

Since neither TEIRESIAS nor PRATT2 is able to find type III patterns, we examined the properties of PGOneIII without comparison. We only show how various windows affect the runtime as the size of the input and the support affect PGOneIII in the same way as PGOneI.

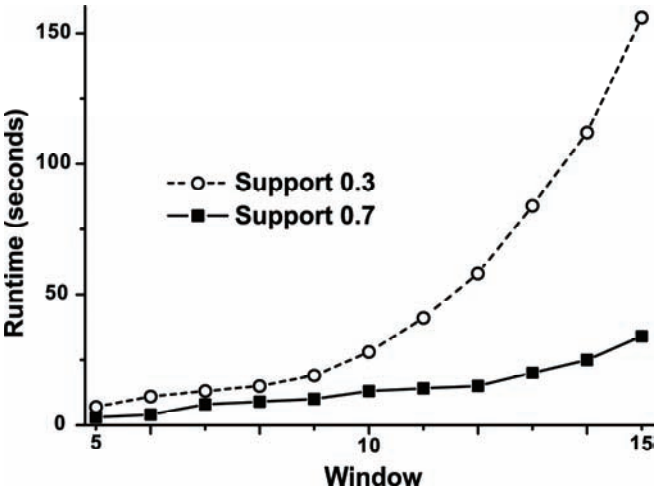


Figure 4. Characteristics of PGOneIII on various windows.

First we examined how the size of the input affects the runtime. The runtime was again linear with the size of the input, when we used a window of size 8, no matter whether patterns with high or low support were searched for.

To understand how runtime changes when PGOneIII searches patterns with different support, we mined the dataset with 200 non-olfactory proteins for patterns having support from 0.1 to 1.0 in a window of 8. Runtime increases very fast when support drops from 1.0 to 0.1. When support is 1.0, most patterns will not grow long since the pattern and its offspring pattern will not be pruned anymore once it fails in any of the proteins in the input. When support drops towards 0, more and more branches of the pattern tree will be searched and eventually PGOneIII will dump every fragment of proteins with a slide window of size 8 as a pattern when support is close to 0.0.

When we increase the window, PGOneIII will search patterns that cover longer fragments in the proteins. We mined the data set of 200 non-olfactory proteins for patterns having a window from 5 to 15. As shown in Figure 4, runtime rose much slower in the search for patterns with high support than with low support.

Classification with PGTwoI, PGTwoII and PGTwoIII

We adapted the pattern growth approach to mine two datasets at the same time and find those patterns (one of the three types) with high support in the predefined positive dataset and low support in the negative one.

We found a series of motifs which distinguish non-olfactory receptors from olfactory receptors. As shown in Table 2, all these patterns are related and indicate that the well-known motif of a cluster of aromatic residues in helix 6 is unique for the non-olfactory receptors (Visiers et al., 2002).

Table 2. Type I Patterns identified by PGTwoI using non-olfactory receptors as positive dataset and olfactory receptors as negative dataset. The maximum number of consecutive wildcards was 3. The minimal support difference was 0.6.

Motif	Support in non-olfactory receptors	Support in olfactory receptors	Support difference
CWxP	0.761	0	0.761
FxxxWxP	0.842	0.003	0.839
FxxCxxP	0.720	0.009	0.711
FxxCW	0.667	0.005	0.662
FxxCWxP	0.667	0	0.667

We also examined which patterns distinguish olfactory from non-olfactory receptors and compared our findings with previous studies on this topic (Mombaerts, 1999). Among other findings, we learned that the sequence of "FSTCSSH", reported to occur in TM6 of a number of olfactory receptors, could be updated as two related patterns "TCxxHxxxV" and "KxxxTCxxH", which show up in more than 80% of olfactory receptors while less than 0.2% of non-olfactory receptors have these. We also found a frequent pattern of "YxxxxxGN" in TM1 which was not included in Mombaerts's study.

Discussion

Reduce search space by pattern growth

A seemingly straight-forward algorithm to find patterns with length L in a dataset of n unaligned protein sequences is to generate and test each pattern candidate. Let us estimate, however, how many pattern candidates are needed to find all type I patterns with a length of 10. Since at each position, both 20 residues and the wildcard x may show up, the total number of pattern candidates to be generated would be $21^{10} \approx 10^{13}$. Moreover, for each pattern candidate, the entire dataset would have to be scanned to record the number of protein sequences having the pattern candidate. Thus, a more sophisticated method is warranted.

In this chapter, we present a pattern growth solution to largely reduce the number of pattern candidates and the cost on scanning the dataset. We grow the pattern by adding elements to its "righthand" side. Once the pattern is no longer frequent, we stop growing it and its extension will not be tested. For example, if pattern AxxDxxxG is not frequent, we will not test patterns such as AxxDxxxGA, AxxDxxxGS and AxxDxxxGx. In this way, we significantly reduce the number of pattern candidates to be tested while all patterns that satisfy the requirement will be reported.

In addition, we avoid scanning the entire dataset for each pattern candidate. Obviously, if one protein doesn't contain a certain pattern, it will not contain the extension of that pattern either. For each pattern that is defined as frequent in the mining process, we create a projected

database to store the index of the protein sequences which contain that pattern. The extensions of that pattern will only be tested on the proteins in the projected database.

Mining various pattern types by pattern growth

The pattern growth approach not only provided us with an efficient solution in pattern discovery but also enabled us to find various types of patterns.

In this chapter we carefully controlled the ways to grow the pattern resulting in three different types of patterns. We expect that the pattern growth approach may find many other pattern types proposed in the future because of the excellent adaptability to various gap constraints.

The three pattern types target protein regions with different levels of conservation. The type I pattern may capture the conserved positions in a region with secondary structure such as functional motifs in a helix or a beta sheet. The type I pattern is also the major outcome of MSA-based approaches.

In some cases, small insertions/deletions may happen in the region with secondary structure yielding type II patterns. For example, proline and/or double glycine are known to introduce kinks in a helix. To maintain the overall structure, an insertion/deletion may then be necessary, of which there are many examples in class A GPCRs. It is more difficult for MSA-based methods to find type II patterns compared to type I patterns.

We also introduce a type III pattern which may target protein fragments without a defined secondary structure. MSA-based methods are not suitable to find type III patterns. Novel as this new type of patterns is for protein sequences, we are faced with the question of their meaning. Type I and type II patterns can be found and interpreted by overlaying homologous protein structures. However, type III patterns may not be found in this way since we cannot overlay regions that do not have a well-defined structure. For example, the third intracellular loop of GPCRs has various lengths in different receptors and is believed to be involved in G protein coupling. Unfortunately, in the only crystal structure of class A GPCRs, bovine rhodopsin, the structure of that loop is not well defined and varies in different studies. Given the fact that we have limited other tools, if any, to find and interpret the biologically important residues in a protein region without a defined secondary structure, our PGOneIII algorithm may become particularly valuable.

Mining two datasets at the same time

We also adapted our pattern growth approach to mine two datasets at the same time to find patterns that distinguish members in the positive dataset from those in the negative one. Of course this can be done by comparing mining results from the two datasets. However, our pattern growth approach provides us with an efficient way to find the complete patterns that satisfy the requirement.

Users may get a better idea about what patterns may be associated with certain properties from mining two datasets. Subsequently, these patterns can be used for classification or to predict new members which share the same properties.

With MSA versus without MSA

Pattern discovery algorithms using aligned protein sequences have been developed for decades. However, ambiguity in an MSA would jeopardize such pattern analysis. In this chapter we proposed a new algorithm to identify patterns directly from unaligned protein sequences. Thus we circumvented the potential problems caused by creating an MSA. From the results, we argue that for a set of proteins with low sequence identity our approach is much better than the approaches based on MSA for several reasons. First, the computation time to build an MSA (hours or even days for a protein family as big as GPCRs) is saved. Second, for proteins with low sequence identity, MSA is not reliable, which will harm the pattern discovery. For example, some functionally important positions may escape detection because they are in or even close to gap regions. On the other hand, when protein sequences are highly homologous, the approaches based on MSA are better because our approach will slow down to search deep in the pattern tree and the computation is expensive and largely meaningless since millions of patterns may be generated from the highly homologous sequences. In addition, MSA becomes reliable for pattern discovery in such cases.

In short, MSA-based approaches may find type I and type II patterns, whereas our approach finds all three types of patterns.

Pattern growth versus TEIRESIAS

TEIRESIAS uses a very different way to efficiently mine type I patterns from one dataset. It first finds frequent short patterns and then glues them together to form a larger pattern. Similar to our approach, TEIRESIAS reports all patterns that satisfy the parameters entered by the user. Although the algorithms are different, PGOneI and TEIRESIAS behave very similar: both succeed in all the tests, the runtime is linear with the size of the input and is affected by the support (or more precisely the number of patterns to find).

Although TEIRESIAS is efficient in mining type I patterns from one dataset, it is not able to find type II and type III patterns.

Pattern growth versus PRATT2

PRATT2 uses a different strategy to mine type I and type II patterns from one dataset. It uses a portion of sequences to create a pattern graph, then tries to find (or optimize) patterns according to its scoring function. Creating and searching a pattern graph is expensive in terms of runtime and memory. That is why it runs slow and failed when the input contained 600 or more proteins in our tests. This failure in processing large datasets significantly reduces the application of PRATT2 in the life sciences, in view of the ever increasing amount of data.

PRATT2 seems to find “best” rather than all patterns that satisfy the requirement because of the nature of its algorithm. Unfortunately, PRATT2 ranks the patterns by its own scoring function and outputs only the top 50 patterns by default. Even if the maximal number of patterns to be reported is set big enough, PRATT2 yields only part of all patterns. This policy looks friendly at first glance but it does not provide all solutions.

CONCLUSIONS

In this chapter, we designed and implemented six algorithms to mine three different pattern types from either one or two datasets using a pattern growth approach. We compared our approach to PRATT2 and TEIRESIAS in efficiency, completeness and the diversity of pattern types. Compared to PRATT2, our approach is faster, capable of processing large datasets and able to identify the type III patterns. Our approach is comparable to TEIRESIAS in discovery of type I patterns but has additional functionality such as mining type II and type III patterns and finding discriminating patterns from two datasets.

References

Agrawal R and Srikant R, (1994) Fast Algorithms for Mining Association Rules, *Proc. 1994 Int’l Conf. Very Large Data Bases (VLDB ’94)*, 487-499.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A and Zygouri C (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 31:400-402.

Baldi P and Chauvin Y (1994) Hidden Markov Models of the G-protein-coupled receptor family. *J Comput Biol* 1:311-336.

Baldi P, Chauvin Y, Hunkapiller T and McClure MA (1994) Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci USA* 91:1059-1063.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C and Eddy SR (2004) The Pfam protein families database. *Nucleic Acids Res* 32(Database issue):D138-141.

Copley RR, Russell RB and Ponting CP (2001) Sialidase-like Asp-boxes: Sequence-similar structures within different protein folds. *Protein Sci* 10:285-292.

Garofalakis M, Rastogi R, and Shim K (2002) Mining sequential patterns with regular expression constraints. *IEEE Transactions on Knowledge and Data Engineering* 14: 530-552

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M and Sigrist CJ (2006) The PROSITE database. *Nucleic Acids Res* 34(Database issue):D227-230.

Jonassen I (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput Appl Biosci* 13:509-522.

Jonassen I, Collins JF and Higgins DG (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci* 4:1587-1595.

Kuipers W, Oliveira L, Vriend G and IJzerman AP (1997) Identification of class-determining residues in G protein-coupled receptors by sequence analysis. *Receptors Channels* 5:159-174.

Lichtarge O, Bourne HR and Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257:342-358.

Lupas AN, Ponting CP and Russell RB (2001) On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* 134:191-203.

Mombaerts P (1999) Seven-transmembrane proteins as odorant and chemosensory receptors. *Science* 286:707-711.

Pei J, Han JW, Mortazavi-Asl B, Wang JY, Pinto H, Chen QM, Dayal U and Hsu MC (2004) Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering* 16:1424-1440.

Pei J, Han JW, and Wang W (2002) Mining Sequential Patterns with Constraints in Large Databases. *Proceedings of the 11th ACM International Conference on Information and Knowledge Management (CIKM'02)*, 18-25

Rigoutsos I and Floratos A (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* 14:55-67.

Russell RB, Saqi MA, Bates PA, Sayle RA and Sternberg MJ (1998) Recognition of analogous and homologous protein folds: Assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng* 11:1-9.

Shigeta R, Cline M, Liu G and Siani-Rose MA (2003) GPCR-GRAPA-LIB: A refined library of hidden Markov Models for annotating GPCRs. *Bioinformatics* 19:667-668.

Visiers I, Ballesteros JA, Weinstein H. Three-dimensional representations of G protein-coupled receptor structures and mechanisms. *Methods Enzymol* 2002;343:329-371.

Ye K, Lameijer EW, Beukers MW and IJzerman AP (2006) A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors. *Proteins* 63:1018-1030.

Chapter 6

Alignment independent phylogeny reconstruction – A cheminformatics approach

Motivation: Phylogeny reconstruction is usually based on multiple sequence alignment. The procedure can be computationally intensive and often requires manual adjustment, which may be particularly difficult for a set of deviating sequences. In cheminformatics, constructing a similarity tree of ligands is usually alignment free. Feature spaces are routine means to convert compounds into binary fingerprints. Then distances among compounds can be obtained and similarity trees are constructed via clustering techniques. In protein sequence analysis several ways to convert protein sequences into binary fingerprints have been proposed for remote homology detection, which is an example of supervised learning. Can we directly use these fingerprints to construct a phylogenetic tree? Can we extract better feature spaces for phylogeny reconstruction in an unsupervised way via sequential pattern mining directly from unaligned protein sequences?

Results: In this chapter, we explored building feature spaces for phylogeny reconstruction either using the *k*-mer method or via sequential pattern mining with additional filtering and combining operations. Satisfactory trees may be built from both approaches when compared with alignment-based methods. We found that when *k* equals 3, the phylogenetic tree built from the *k*-mer fingerprints is as good as one of the alignment based methods, in which PAM and Neighborhood joining are used for computing distance and constructing a tree, respectively (NJ-PAM). As for the sequential pattern mining approach, the quality of the obtained phylogenetic tree is better than the tree built from NJ-PAM, when we set the support value to 10% and used only maximum patterns as descriptors.

Introduction

Construction of a phylogenetic tree is one of the fundamentals in bioinformatics. It describes how a protein family might have been evolved during evolution. Sequences that are the most closely related can be identified if they occupy neighboring branches on a tree. Most of the approaches for phylogenetic analysis are strongly based on multiple sequence alignment (Fink, 1986). Similar or even identical parts between pairs of sequences are first identified and aligned. Then the evolutionary distances among the sequences are calculated according to one of the models. Finally such a distance matrix is used to construct the phylogenetic tree (Ma et al., 2007; Retief, 2000; Tamura et al., 2007). Alignment-dependent phylogeny construction, however, inherits the pitfalls of creating an alignment. First, it is a time-consuming process to align hundreds or thousands of sequences. Second, an alignment may be tuned by various parameters such as gap opening and extension penalties which may affect the final phylogeny. Third, the distance between two sequences largely emphasizes the well-aligned regions. Last but not least, as sequences become more and more divergent through a long period of evolutionary history, it is more difficult to obtain an accurate alignment and then reconstruct a reliable phylogenetic tree. Having all these issues in mind, can we reconstruct a phylogenetic tree without aligning sequences?

In cheminformatics, clustering small molecules does not rely on aligning thousands of compounds in 2D and overlaying their millions of potential conformations. A so-called *fingerprints* technique is often used instead. Fingerprints are an abstract representation of certain structural properties of a molecule. If we decide to use n structural features to represent all compounds, then each of these will be coded as a binary string with n bits. If certain structural features are present or absent in the compound, the corresponding bits will be set to one or zero, respectively. State-of-the-art fingerprint methods include dictionary-based (Barnard and Downs, 1997) and layered atom environment fingerprints (Bender et al., 2004). If we accept that the proportion of structural features shared by two molecules is a reasonable similarity measure of these two molecules, we may compare the bits which are set as one between the two molecules. Many distance measures, such as matching, Cosine, Euclid, Tanimoto and Russell-Rao, have been proposed to compute a distance from two binary fingerprints (Willett et al., 1998).

If we build a multiple sequence alignment only for the purpose of calculating distances among sequences for phylogeny construction, can we instead obtain those distance values without aligning sequences following the principle of measuring the similarity of two compounds? We notice that when we calculate the distance between two aligned sequences, the more consensuses observed the more similar the sequences are. If we present the consensus parts as features, the representations of sequences in the feature space may be used to measure sequence similarity. If two sequences share more features, they are more similar. Several

feature spaces for protein sequences have already been proposed and can be classified as two categories: predefined feature spaces and dataset-specific feature spaces.

Leslie and coworkers mapped protein sequences to *k*-mer feature space (Leslie et al., 2004). The *k*-mer feature space has all possible amino acid combinations of a fixed length *k*. When *k* equals 1 or 2, almost all features may show up in every protein sequence. Thus larger *k* values are desired. When *k* equals 3 or 4, a protein sequence is represented as a vector of 8,000 (20^3) or 160,000 (20^4) features, respectively.

The *k*-mer approach provides constant collections of features which will not differ when we analyze various protein families. It is equally possible to extract conserved patterns as dataset-specific features from a given protein family using alignment-independent pattern discovery algorithms such as TEIRESIAS (Darzentas et al., 2005; Dong et al., 2006; Rigoutsos et al., 1999) as well as our own pattern growth program (Chapter 5). Rigoutsos et al. considered protein sequences as a language and applied TEIRESIAS to collect the "words" (Seqlets). They aimed at finding as many of the words as possible for high or even complete coverage of the sequence space. The *k*-mer method and TEIRESIAS have been applied in remote homology detection, a supervised learning method (Ben-Hur and Brutlag, 2003; Darzentas et al., 2005; Dietmann et al., 2002; Dong et al., 2006; Leslie et al., 2004; Saigo et al., 2004; Zhang et al., 2005). In this case, training datasets were included for feature selection or feature weighting, and a Support Vector Machine (SVM) was used for classification. However, in this project, which essentially is an unsupervised learning procedure, a training dataset to manipulate the features is not available, since we aim at phylogeny reconstruction from unaligned protein sequences. Are we able to construct a reliable phylogenetic tree using the complete feature spaces of either the *k*-mer method or pattern discovery approach? For the latter there is a further notion though. For example, when we mine class A G protein-coupled receptors, a protein family characterized by its seven transmembrane helical domains, both pattern NxxNPxxY and its sub-pattern NPxxY are present. The latter one is more general and covers more sequences than the former one. In computer science, the patterns that have no further specific ones are called *maximal*. In this case, pattern NxxNPxxY is maximal. It is not clear, however, how non-maximal patterns affect the similarity measure among sequences. More importantly, if we consider motifs as the words, the sequential order of these words may also carry essential biological meaning. In English, we prefer to start a sentence with a subject, then the verb and we put the object at the end. Protein sequences must also have their patterns in the correct order to fold correctly and function properly. For example, in class A G protein-coupled receptors, the patterns DRY, WLP and NPxxY characterize the helices 3, 6 and 7, respectively, and they are important for the receptors to maintain their seven-transmembrane helical architecture and to switch to an active conformation upon agonist binding. In this chapter we also mine such "combined patterns" (sequential order of patterns) and compare them with normal patterns in phylogeny reconstruction.

Methods

Mining patterns from unaligned protein sequences

Previously we developed an efficient, versatile and scalable pattern growth approach to mine frequent patterns from unaligned protein sequences (Chapter 5). From one dataset, we can identify three different types of patterns which target at the protein regions with different levels of conservation. Among these three types, Type I is the most restricted one and provides information about well-structured protein regions. It yields patterns with a defined number of wildcards such as in AxxT, where x is the wildcard. Although the algorithms are different, the results of the algorithm that mines type I patterns are equivalent to TEIRESIAS. The in-house implementation of the pattern growth approach facilitates additional operations on the discovered patterns.

We start with mining type I patterns with the same parameter setting as in our previous study (Chapter 5), and follow an approach represented in Figure 1 and detailed in the paragraphs below.

Filtering operator

In all patterns discovered in the mining step (Section 2.1), a pattern, such as NxxNPxxY, co-exists with its less specified forms (NxxNP, NPxxY and NxxNxxxY) that also satisfy the mining conditions. If the pattern NxxNPxxY does not have a more specific form, we call it *maximal*. In this filtering step, we remove all non-maximal patterns and keep only maximal ones for the next step.

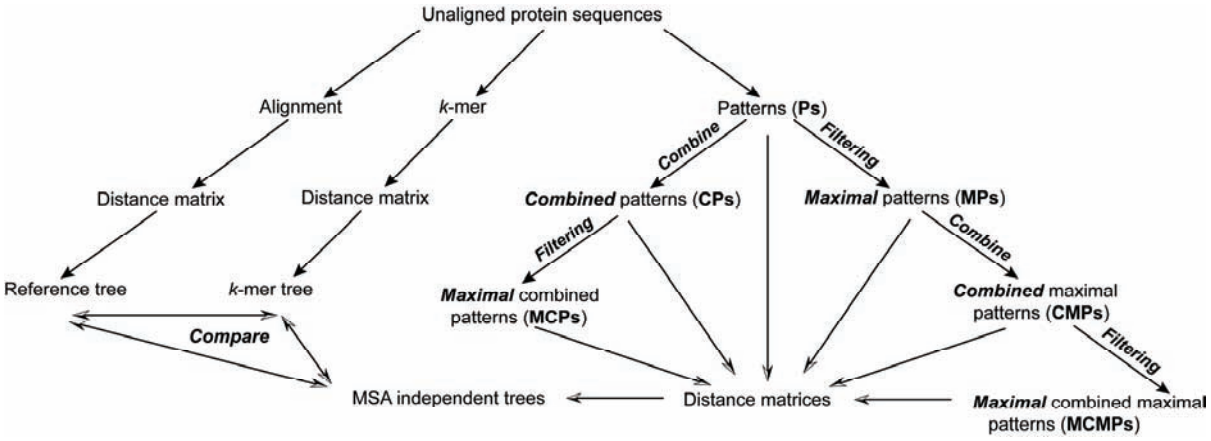


Figure 1. Schematic flowchart of exploring and comparing various feature spaces for phylogeny reconstruction. A reference tree is built via traditional alignment based method; We also use *k*-mer (k = 3, and 8,000 features) to construct a *k*-mer tree for comparison. From unaligned protein sequences, we first identified frequent type I patterns given certain support values. After that we explored various combinations of combine and filtering operations to derive different feature spaces in an unsupervised way.

Combine operator

Not only the presence of motifs but also their relative locations in the sequence may carry essential information about their function. For example, the pattern DRY is always before NPxxY in class A G protein-coupled receptors. Thus we also mine the frequent sequential order of patterns within a given length of the window and call them a *combined* pattern.

The input of the algorithm is a dataset *S* which consists of a series of patterns per protein data structure. Each basic structure describes one protein sequence and includes the patterns and their locations. The input also contains a minimum support threshold *min_support*, the minimal number of patterns in the combined patterns to be reported *min_pattern_num* and the maximum distance between the locations of the first pattern and the last pattern, *window*.

The output of the algorithm consists of all combined patterns that have support larger than or equal to *min_support*, and that satisfy the other input parameters. For each *combined* pattern, its support and the list of proteins that contain it are also given.

The algorithm is as follows, see below. Here a is a combined pattern and S_a is the so-called projected database that contains all sequences that satisfy the combined pattern a, where the last element of each occurrence of a is marked (the so-called a-locations). Note that $|S_a|$, the number of sequences in S_a, is the support of a. Let *Length* (a) be the number of patterns occurring in sequential order in combined pattern a.

```
CombinePattern (a, S_a)
    if |S_a| ≥ min_support then
        if Length (a) ≥ min_pattern_num then
            report a, |S_a|
        for each frequent pattern b do
            a' := a with b appended to it
            CombinePattern (a', S_a')
```

The computation of $S_{a'}$ from S_a requires, for each a-location, the check whether or not the newly appended pattern b is on its right-hand side within the window calculated from the start of a. In case of finding pattern b, we update the a'-location with the location of pattern b. Otherwise we remove the sequence from S_a.

The main call is *CombinePattern*(Λ, S_Λ), where Λ is the empty combined pattern. This call creates a projected database that marks all occurrences of a single pattern b, reports all frequent combined patterns that begin with this b, and then proceeds to the next pattern.

k-mer

The *k*-mer feature space contains all possible amino acid combinations of a fixed length *k*. Since all protein sequences probably contain every instance of a 1-mer, we examined the situations where *k* values varied from 2 to 8. This yielded a series of feature spaces, the sizes

of which ranged from 20^2 (400) to 20^8 (25,600,000,000). In this case, each protein sequence is represented as a binary fingerprint that contain 20^k bits, with dozens or hundreds of bits set to 1 if the corresponding features show up in the protein.

Distance matrices and tree construction

Using either the *k*-mer method or the frequent patterns identified in sequential pattern mining, we may represent each protein as a binary fingerprint, where a particular bit is set to 1 if the corresponding pattern appears in the sequence. Then we may calculate the distance between two proteins by comparing their binary strings.

We chose Russell-Rao distance to calculate the distance between two proteins which are represented by two binary strings based on the observation that more common patterns are shared by the two protein sequences when there is more similarity between them. Thus only the bits that are set as one in both proteins are important. We compared the feature spaces of the *k*-mer approach (*k*=2 - 8) and the various feature spaces derived from pattern discovery with lowest support values possible.

Support *a* is the count of bits on in the fingerprint of protein A but not in the fingerprint of protein B;

b is the count of bits on in B but not in A;

c is the count of the bits on in both A and B;

d is the count of the bits off in both A and B.

Then Russell-Rao distance is

$$\frac{a+b+d}{a+b+c+d}$$

After we have calculated the Russell-Rao distance between all pairs of protein sequences, we obtain a distance matrix. Then we use the neighbor-joining method as implemented in the software package PHYLIP to reconstruct the phylogenetic tree.

Reference tree and control trees

To quantitatively estimate the phylogenetic trees reconstructed from fingerprints, we compared them with a reference tree that is built from a traditional alignment-based method using MEGA4 (Tamura et al., 2007). We first used ClustalW (Ma et al., 2007) to build an alignment with the following simple parameters: gap opening and extension penalty for pairwise alignment are set to 10 and 0.1, respectively, while in the multiple alignment those parameters are set to 10 and 0.2. The protein weight matrix was set to identity and we turned off all other features such as residue-specific penalties and hydrophilic penalties. Then we set

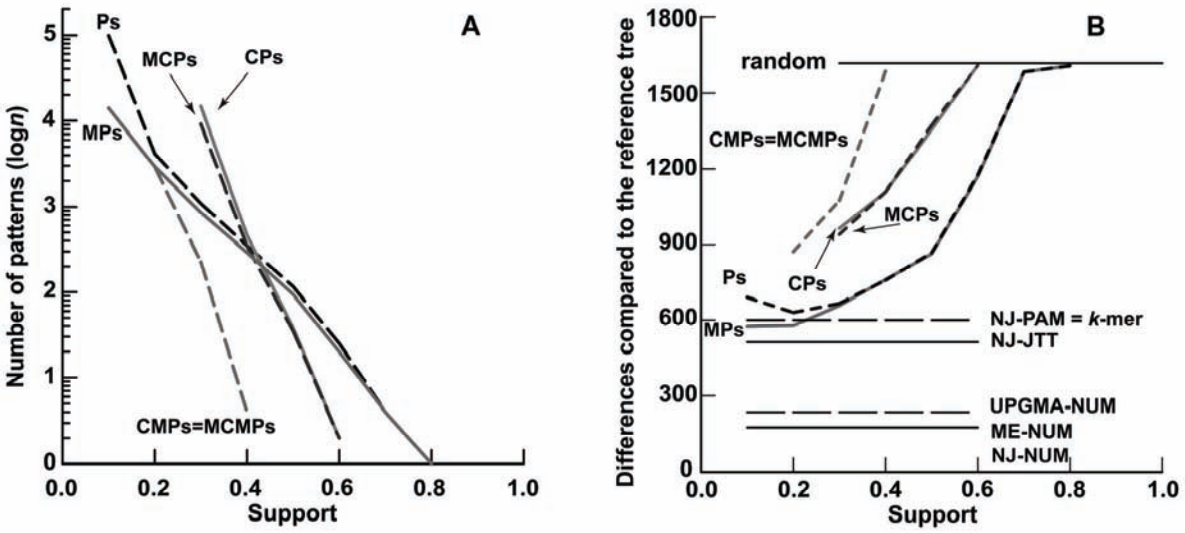


Figure 2. Comparison of the sizes of feature spaces and their qualities in constructing a phylogenetic tree. We set the number of differences as distance measure and used the neighbor-joining method to reconstruct the reference tree (*x*-axis). We kept distance measure as the number of differences but also used either UPGMA or Minimum Evolution (ME) to build two control trees (UPGMA-NUM and ME-NUM). We further built two control trees using the same neighbor-joining tree constructing method, but two different distance measures, either PAM or JTT matrix (NJ-PAM and NJ-JTT). A random control tree was computed by the neighbor-joining method from a random distance matrix. A tree built from *k*-mer (k = 3) feature space is also included as a control. A) The number of patterns in a given feature space with certain support values. B) the qualities of the constructed phylogenetic trees compared with the reference tree built via the alignment-based method. Note that the smaller values in the *y*-axis are, the more similar the trees are compared to the reference tree. Ps = all patterns; MPs = maximal patterns; CMPs = combined maximal patterns; MCMPs = maximal combined maximal patterns; CPs = combined patterns; MCPs = maximal combined patterns.

the number of differences as distance measure and used the neighbor-joining method to build the phylogenetic tree (NJ-NUM).

While we set the NJ-NUM tree as the golden standard, we also constructed four alignment-based control trees to investigate the qualities of fingerprints-based trees. Firstly we changed the distance measure to either PAM or JTT to obtain NJ-PAM or NJ-JTT, respectively. We also obtained UPGMA-NUM and ME-NUM trees using UPGMA or Minimum Evolution (ME) as tree construction algorithms. A random control tree was computed by the neighbor-joining method from a random distance matrix.

Tree comparison

We scored each tree built from fingerprints by comparing it with the reference tree. Treedist implemented in the software package PHYLIP (Retief, 2000) was used to compute the distance between two trees.

We used Symmetric difference (Robinson and Foulds, 1981) which does not use information about branch length but only topological differences. Please notice that the smaller the score is (*y*-axis in Figure 2B), the more similar the tree is to the reference tree.

Dataset

G protein-coupled receptors (GPCRs) are important drug targets. They are membrane-bound, serpentine-like, proteins with 7 transmembrane, alpha-helical, domains. A subdivision in three classes has been proposed of which class A (rhodopsin-like) is by far the largest.

Protein sequences of class A GPCRs were extracted from the GPCRDB (March 2005 release (9.0); http://www.gpcr.org/7tm/). We first removed orphan receptors and olfactory receptors. Then protein sequences without function annotation in the sequence name were removed. This yielded a total of 811 protein sequences.

As shown in Figure 1, we first discovered frequent patterns from unaligned protein sequences. Then we examined various combinations of combine and filtering operators in order to achieve better fingerprints that yield more similar phylogenetic trees compared to the reference tree. In addition to the alignment-based tree (Retief, 2000), we also compared our method with the trees built using the *k*-mer method for collecting fingerprints. We also evaluated the impact of different distance measures on the quality of phylogeny reconstruction.

Results

We aim to discover unsupervised procedures to derive fingerprints from unaligned protein sequences with good qualities in terms of clustering protein sequences as similar as possible to the reference phylogenetic tree built from a sequence alignment.

We explored various feature spaces for the construction of a phylogenetic tree in unsupervised learning. First we investigated dataset-specific feature spaces. We applied sequential pattern mining to unaligned protein sequences and then explored various combinations of filtering and combining operations to manipulate the patterns in an unsupervised way. Then we considered the presence or absence of certain patterns as the fingerprints of the protein sequences. The Russell-Rao distance measure converts fingerprints into distances among protein sequences. Phylogenetic trees were constructed and compared with a reference tree built from a sequence alignment. Secondly, we examined the qualities of trees using the *k*-mer method for feature spaces with *k* values varying from 2 to 8. Finally we compared the qualities of the trees built from either the *k*-mer method or pattern discovery approach.

Number of dataset-specific patterns in each step

- ***Ps*** As a start, we used the pattern growth algorithm developed in our previous study (Chapter 5) to identify frequent patterns by varying support from 1.0 to 0.1. The

support values constrain the adamancy of patterns, and represent the proportion of protein sequences that contain the identified patterns. As shown in Figure 2A (Ps), no patterns satisfying the input parameters were discovered when the support value was set to 0.9, while only one was found when the support value was set to 0.8. When we decreased the support values, the number of patterns satisfying the input parameters increased almost exponentially. When support was 0.1, 99,166 patterns were identified. When we applied various combinations of combine and filtering operations on the patterns, we obtained different pattern pools.

- ***MPs*** When we applied the filtering operation to the patterns (Ps), the numbers of removed non-maximal patterns gradually increased when support values decreased.
- ***CMPs*** When we combined the maximal patterns (MPs), there were no combined patterns if support values were from 0.5 to 0.8. On the other hand, when we searched for combined patterns with support value 0.1, the computation consumed too much memory to complete due to the large number of maximal patterns (13,985).
- ***MCMPs*** This additional filtering operation did not remove any patterns. Thus MCMPs equal CMPs.
- ***CPs*** If we directly applied the combine operation to all frequent patterns, we had two combined patterns when support was 0.6. The number of combined patterns (CPs) increased exponentially when the support value decreased from 0.6 to 0.3, below which it was too computationally expensive to finish.
- ***MCPs*** This additional filtering operation removed a negligible number of combined patterns (CPs).

Support values and various combinations of filtering and combining operations

For each protein, we used identified patterns as fingerprints to represent it. If there are n patterns, each protein is a binary string of n bits in which certain bits are set as one if the corresponding patterns are present in this protein. We then used the Russell-Rao measure (Section 2.5) to calculate distances among protein sequences. The Neighbor-joining method was used to construct a phylogenetic tree. In order to examine which pattern collection carries better discriminative fingerprints and which support value should be used to harvest frequent patterns, we compared the qualities of trees built from fingerprints with the reference tree (Section 2.6).

As shown in Figure 2B, the top horizontal control line (‘random’) indicates the quality of a tree built from a randomly generated distance matrix while the quality of the reference tree, set as the golden standard, is represented by the x-axis.

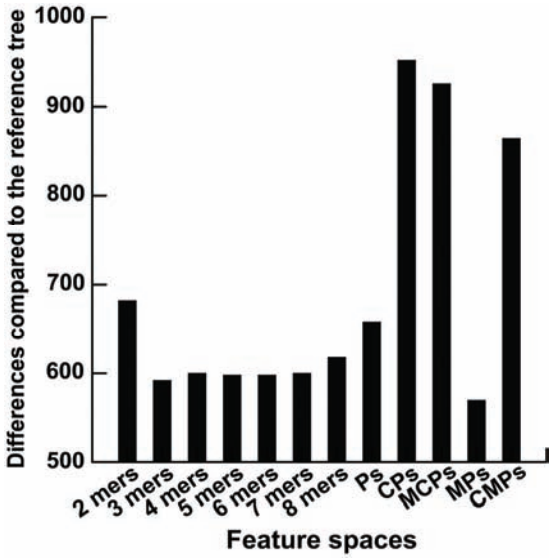


Figure 3. Evaluation of the *k*-mer method (*k* values varied from 2 to 8) and comparison with feature spaces derived from sequential pattern mining. For the latter, the support values were set as the lowest values possible: Ps (0.1); MPs (0.1); CMPs (0.2); CPs and MCPs (0.3). Note that the smaller values on the *y*-axis are, the more similar the trees are compared to the reference tree. Ps = all patterns; MPs = maximal patterns; CMPs = combined maximal patterns; MCMPs = maximal combined maximal patterns; CPs = combined patterns; MCPs = maximal combined patterns.

When we used the same tree reconstruction algorithm (NJ) but different distance measures of NUM, we obtained trees that have a score around 600 (Figure 2B). Notice that the tree built on *k*-mer fingerprints (*k* = 3) also has a score in this range. However, if we used the same distance measure but different tree reconstruction algorithms (UPGMA or Minimum Evolution), we obtained very similar trees to the reference tree.

As shown in Figure 2B the qualities of the trees increased dramatically from "random" towards the quality of the reference tree when we decreased the support value. It is interesting to notice that when support reaches a value of 0.1, the quality of the tree built on all patterns (Ps) decreased. This may be due to the fact that the number of noise patterns unrelated to correct classification increases faster than the number of discriminative ones at such low support values.

When we compared the qualities of the trees built from pattern collections before and after a filtering operation (Ps versus MPs), we found that maximal patterns (MPs) always performed significantly better than just using all patterns (Ps) as fingerprints.

The combine operation, however, had limited value in improving tree qualities. We could only obtain combined patterns within a narrow range of support values (Figure 2B). When the support value is large, the number of patterns is too few to form frequent combined patterns. On the other hand, when the support is small, there are too many candidate patterns to form combined patterns. In that case, it was not longer feasible to finish the calculation.

k-mer feature space

In addition to the dataset-specific feature spaces, we also examined the qualities of the tree using *k*-mer (k values varied from 2 to 8) fingerprints. As shown in Figure 3, even when *k* equals 2, the constructed phylogenetic tree resembles the reference tree to some extent. When *k* equals 3, a tree built from *k*-mer feature space is as good as one of the alignment-based

phylogeny reconstruction procedures (Figure 2B). However, when we increased *k* values even further, the tree qualities became slightly worse (Figure 3). We compared *k*-mer feature spaces with the ones derived from pattern discovery with the lowest possible support values. Based on the current experimental setting, the best solution so far on the dataset of G protein-coupled receptors is to first identify frequent patterns and then use only maximal patterns as fingerprints for phylogeny reconstruction.

We also examined the impact of different distance measures (Tanimoto, Russell-Rao, Cosine, Dice, Kulczynski, Manhattan and Rogers-Tanimoto) on the qualities of phylogeny reconstruction. The results showed that using various distance measures has little effect on the tree qualities (data not shown).

Discussion

In this chapter we explored various feature spaces from either the *k*-mer method or pattern discovery in order to reconstruct a phylogenetic tree. For the dataset of G protein-coupled receptors, the best feature space is obtained by mining frequent patterns and then removing non-maximal ones. We had to use small support values to obtain a large amount of maximum patterns representing a protein sequence in high dimensional space. Combining patterns seems promising since the quality of the tree generated increased dramatically when support values decreased. However, no definitive conclusion can be drawn at this moment since the calculation is computationally too expensive when support values dropped towards 0.1. We may further optimize the mining process to speed up and improve pattern quality. For example, we may increase a parameter when mining patterns, e.g., the maximum number of consecutive gaps. This will yield patterns with longer maximal gaps and of course more and longer patterns to capture the properties of larger fragments in the protein sequences.

Currently we are using exact matches during the mining process. This means that the two patterns DRY and ERY are treated as two distinct patterns. Since residues D and E are comparable, i.e., both are negatively charged, one may combine them as one pattern (D/E)RY. One of the solutions to solve this is to define similar residues as equivalent so that we can combine their instances during mining. Another solution is to include a standard amino acid substitution matrix such as BLOSUM62 into sequential pattern mining. Thus similar patterns will be combined if their scores to the center pattern are above a given cutoff.

When we manually examined the patterns (Ps), we found a large number of patterns composed of hydrophobic residues such as LxxxLxL and LxxxLxxL when we mined the dataset of G protein-coupled receptors. Since G protein-coupled receptors contain a large portion of residues that non-specifically interact with the fatty cell membrane, those “hydrophobic” patterns may not carry essential information to discriminate protein sequences. Hence, we may remove these patterns which are not statistically significant compared to background residue frequencies.

In this chapter, we mined the dataset only once when we used sequential pattern mining to build a feature space. All subsequent operations were performed from this starting collection. We may also use a divide-and-conquer strategy to recursively mine the feature space and divide the dataset into smaller sets. For example, we may start with relatively high support values (as high as 0.5) to get a few frequent patterns. Then we may use these patterns to divide the dataset into several groups. After that, we may use the same support value to mine patterns for each group and use those newly obtained patterns to divide the current group further. This will yield a tree if we recursively repeat such mining and dividing operations.

Conclusion

In this chapter we investigated alignment-independent phylogeny reconstruction. Similar to cheminformatics approaches, we used a so-called *fingerprints* technique to represent each protein sequence as a binary fingerprint. We were able to obtain phylogenetic trees that were comparable in quality to one of the alignment-based approaches.

We examined the qualities of these trees in more detail. In the *k*-mer method with *k*=3, the quality of the tree was better than for other *k* values. In the sequential pattern mining approach we explored various combinations of filtering and combining operations to manipulate the patterns in an unsupervised way. When we used maximum patterns (MPs) as fingerprints for phylogeny reconstruction, the quality of the tree was better than those built with the *k*-mer method.

To conclude, tools and approaches from the field of cheminformatics can be used and tuned to resolve issues of protein clustering that are usually regarded as typical for bioinformatics. The principle of clustering compounds by comparing their fingerprints, as in cheminformatics, allowed an alignment-independent phylogeny reconstruction, which may be useful in protein families with low sequence identity.

References

Barnard, J.M. and Downs, G.M. (1997) Chemical fragment generation and clustering software, Journal of Chemical Information and Computer Sciences, 37, 141-142.

Ben-Hur, A. and Brutlag, D. (2003) Remote homology detection: A motif based approach, Bioinformatics, 19 Suppl 1, i26-33.

Bender, A., Mussa, H.Y., Glen, R.C. and Reiling, S. (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance, Journal of Chemical Information and Computer Sciences, 44, 1708-1718.

Darzentas, N., Rigoutsos, I. and Ouzounis, C.A. (2005) Sensitive detection of sequence similarity using combinatorial pattern discovery: A challenging study of two distantly related protein families, Proteins, 61, 926-937.

Dietmann, S., Fernandez-Fuentes, N. and Holm, L. (2002) Automated detection of remote homology, Curr Opin Struct Biol, 12, 362-367.

Dong, Q.W., Wang, X.L. and Lin, L. (2006) Application of latent semantic analysis to protein remote homology detection, Bioinformatics, 22, 285-290.

Fink, W.L. (1986) Microcomputers and Phylogenetic Analysis, Science, 234, 1135-1139.

Leslie, C.S., Eskin, E., Cohen, A., Weston, J. and Noble, W.S. (2004) Mismatch string kernels for discriminative protein classification, Bioinformatics, 20, 467-476.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2007) ClustalW and ClustalX version 2.0, Bioinformatics, 23, 2947-2948

Retief, J.D. (2000) Phylogenetic analysis using PHYLIP, Methods Mol Biol, 132, 243-258.

Rigoutsos, I., Floratos, A., Ouzounis, C., Gao, Y. and Parida, L. (1999) Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins, Proteins, 37, 264-277.

Robinson, D.F. and Foulds, L.R. (1981) Comparison of phylogenetic trees, Mathematical Biosciences, 53, 131-147.

Saigo, H., Vert, J.P., Ueda, N. and Akutsu, T. (2004) Protein homology detection using string alignment kernels, Bioinformatics, 20, 1682-1689.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, Mol Biol Evol, 24, 1596-1599.

Willett, P., Barnard, J.M. and Downs, G.M. (1998) Chemical similarity searching, Journal of Chemical Information and Computer Sciences, 38, 983-996.

Ye, K., Kosters, W.A. and IJzerman, A.P. (2007) An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences, Bioinformatics, 23, 687-693.

Zhang, Z., Kochhar, S. and Grigorov, M.G. (2005) Descriptor-based protein remote homology identification, Protein Sci, 14, 431-444.

Chapter 7

Conclusions and perspectives

Conclusions from the thesis

It is fair to say that the research discipline of bioinformatics largely emerged from sequence analysis. The most important components of life, DNA, RNA and protein molecules are linear sequences by nature. Deciphering the language of life has intrigued numerous researchers for the past half century. The methods of aligning sequences, building phylogenetic trees etc. were invented decades ago. In this thesis, we show that novel algorithms are still in demand because discovery of all hidden information from protein sequences is by no means finished yet.

It is well accepted in sequence analysis that conserved residues are likely to be functionally important. Indeed, many early approaches select functional sites by simply picking the most conserved positions in a given multiple sequence alignment. Since vast amounts of sequence data have become available, sequence comparison between paralogous and orthologous proteins is performed routinely in order to identify specificity residues that account for functional differences.

We designed two series of methods, information-entropy-based and Machine-Learning based, to identify specificity sites in a protein family. From a given classification or a phylogenetic tree, information entropy was used to identify residues conserved within subfamilies but divergent among them. We further developed multi-RELIEF based on RELIEF, a state-of-the-art Machine-Learning technique for feature weighting, for the identification of specificity residues. Incorporating structural information into the learning process improved the prediction for specificity of interaction with small molecules.

Protein motif discovery is tightly connected to multiple sequence alignment. Here we proposed a complete solution to find conserved sites from protein sequences through a new algorithm that directly identifies frequent biologically meaningful patterns from unaligned sequences. Six algorithms were designed and implemented to mine three different pattern types from either one or two datasets using a pattern growth approach. Our approach is better than two state-of-the-art algorithms, PRATT2 and TEIRESIAS, in efficiency, completeness and the diversity of pattern types. Phylogeny reconstruction has always been thought to be alignment dependent. However, we demonstrated that a cheminformatics-like approach can yield a decent phylogenetic tree. We just need to extract patterns from unaligned protein sequences and then use those patterns as fingerprints to compute a distance matrix from which a tree can be easily built.

Perspectives

TEA as a feature visualization tool

TEA (Chapter 2 and 3) is successful in protein sequence analysis when either a classification or a tree-like structure is known or given. If we consider each protein as an entity and each position as a feature, TEA is actually a feature visualization tool and we may apply it to many other fields such as cheminformatics. The features shared by most of the entities will generally be at the lower left corner while those that correlate with classification or tree splitting will appear at the upper left corner of the TEA plot. When we visualize the movement of features on the plots by tracing from the root to the branches of a tree, each feature has its own properties in terms of height in the plot (*y*-coordinate) and variations of speed when we travel from one level to another in the tree on the *x*-axis. The movements of features reflect their relative contribution to the tree splits at all levels.

TEA in cheminformatics

TEA originates from protein sequence analysis and hence belongs to bioinformatics. However, selection and visualizing features of compounds for a given classification or tree-like structure may also be valuable. For example, if we have a compound library classified into several groups or into a tree-like structure, we may use TEA to visualize features shared by all compounds and those that discriminate between different classes.

In addition, we may feed TEA with both protein sequences and a compound library at the same time to find residues that are important for the binding of certain classes of compounds and what kind of properties of compounds contribute to their binding to particular proteins.

Mining frequent patterns in biological sequences

Most state-of-the-art pattern discovery and functional site prediction algorithms work by first building a multiple sequence alignment (Mulder and Apweiler, 2007; Wu et al., 2006; Ye et al., 2007; Ye et al., 2006). Currently our algorithms find frequent biologically meaningful patterns such as PROSITE-like patterns directly from unaligned protein sequences as well as discriminating patterns between two groups of sequences. We may adapt our algorithms to find frequent patterns in DNA/RNA sequences. Either the patterns showing up in a certain portion of sequences (a set of co-regulated genes, chromosomes in one species, one particular chromosome such as the Y chromosome or even entire genomes from different species) or those present in multiple locations in one sequence may be defined as frequent.

Mining unique-m probes in genomes

Currently microarray analysis is very important to understand gene regulation. It uses a small fragment of DNA to fish for particular RNA molecules in the test sample. The probe will capture those RNA molecules that contain a perfect reverse-complement counterpart.

However, sequences with a couple of mismatches with the probe may also be caught, albeit with less intensity. We may consider a short DNA fragment as a poor probe if it has quite a few approximate matches in the genome. Those approximate mismatches can be very confounding since it is not longer sure which RNA molecules are in the sample. As the number of mismatches increases, the possibility of RNA being fished by the probe drops dramatically. Thus we may use the number of mismatches as a rough measure for the probability of fishing approximate RNA by a given probe. We define one DNA fragment as a unique-m probe if its reverse complement matches one spot on the genome perfectly and all the other approximate matches have more than *m* mismatches. In this way we would be able to design better DNA probes in an unprecedented way. Similarly, we may also use these unique-m strings to design RNAi probes that have reduced off-target effects (Moffat et al., 2007; Rual et al., 2007).

Mining forbidden patterns in biological sequences

Frequent patterns are functionally important and are under positive selection. We may also learn what kind of sequences are under negative selection so that they are forbidden in biological sequences, particularly in genomes. Finding forbidden patterns may provide a different angle to interpret natural selection and how life evolved. In addition, they are considered useful for a variety of basic science and application scenarios, including drug target identification, pesticide development, environmental monitoring and forensics (Hampikian and Andersen, 2007).

We can tune the sequential pattern mining algorithms we designed and implemented to find forbidden patterns in a collection of protein sequences as well as several genomes.

Hunting RNA nanomachines in genomes

As we know mRNA is transcribed from DNA and then translated into protein. Textbook biochemistry teaches that most transcribed regions in genomes are protein-coding since this strategy is efficient in terms of building material and energy. Therefore 98.5% of the genome is often referred to as “junk” DNA since it does not code for proteins. The recent Encyclopedia of DNA Elements (ENCODE) pilot project, however, surprised the research community by showing that the majority of the human genome is associated with at least one primary transcript. Most RNA molecules are not produced for the synthesis of proteins. If some of these RNA molecules perform certain biological functions, they will probably fold to stabilize and exhibit particular 3D conformations to perform their biological activities.

As a single chain molecule, RNA will first form hairpins and then several hairpins fold into more complicated 3D structures. The 16S RNA and transfer RNA (tRNA) are two examples of such RNA nanomachines (Figure 1). Hence, to search for RNA nanomachines in genomes, we may start with mining statistically significant hairpins. If the density of hairpins in a small region is larger than the tRNA or 16S RNA coding regions, the DNA sequence in this region

may code for an RNA nanomachine. This hypothesis may be checked from genome annotation or tested in “wet” experiments.

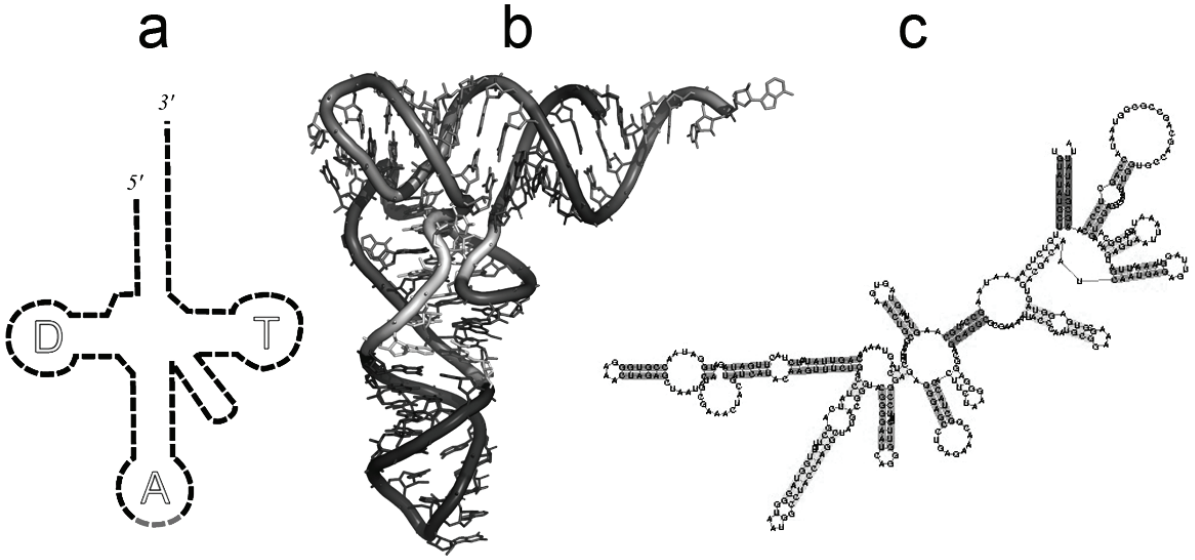


Figure 1. a) transfer RNA; b) structure of transfer RNA; c) 16S RNA

Mining frequent modules from biological networks via sequential pattern mining

Biological molecules do not function in isolation. They are linked via interaction, chemical reaction and modulation. The entire biological network, however, is too complicated to be analyzed. Sophisticated mining algorithms are necessary to extract general rules or basic functional subnetworks to gain insight into the network. In this way we may learn the properties of a network as well as means to intervene with it. Again, sequential pattern mining algorithms can be adapted now, to mine frequent functional modules (subnetworks).

Bioinformatics and computer science

When we want to solve a biological problem with computational approaches, we may start with searching for appropriate algorithms in computer science. Often, as we have learned, such algorithms need to be further tuned to the special need of the biological problem. Despite this, the advantage of using algorithms developed in computer science is enormous. First, these algorithms are very efficient, which is very important in the era of data explosion. Secondly, it is very easy to incorporate additional constraints to the algorithms since they are well studied and characterized. Last but not least, understanding certain algorithms for a given biological problem may also be beneficial to tackle related but non-identical biological problems.

It is equally possible to develop new algorithms by thoroughly studying biological problems. TEA and TEA-O in this thesis are two such examples. This goal-oriented algorithm development is practical and works without a tuning step.

Software user, maker and algorithm designer and implementer

Researchers in either cheminformatics or bioinformatics can be generally classified into two categories. In the first category, researchers tend to use existing software packages to solve particular problems in their experimental work. I was in this category during the first year of my PhD project when I used InsightII to build homology models of adenosine receptors and AutoDock to investigate the potential interactions between various ligands and receptors. During the sequence analysis of adenosine receptors, I noticed that most of the residues in the ligand binding site of adenosine receptors are conserved within adenosine receptors but divergent in other class A G protein-coupled receptors. To my knowledge of that time there was no algorithm available to analyze such differences for a whole family of proteins. Thus encouraged by Prof. IJzerman and with the help of Eric-Wubbo Lameijer, PhD student and a senior programmer in our group, I started my first C/C++ project.

This brings me to the other category: program maker rather than user. Such a transition proved to be fruitful. We invented TEA and later modified a sequential pattern mining program for protein sequence analysis. With my biological background and equipped with programming skills, I enjoyed collaborating with computer scientists which yielded two Bioinformatics papers co-authored with Walter Kosters and Elena Marchiori, respectively.

The software makers may be further divided into algorithm designers and implementers. In many cases, the latter one receives much more attention compared to the former. For example, most citations referring to the methodologies of multiple sequence alignment and phylogeny reconstruction were from peer researchers who are also exploring novel approaches. The software package MEGA3 was published in 2004. Many popular alignment and phylogeny reconstruction algorithms were implemented in it and a nice mouse-windows interface was provided, while most other packages at that time included only one method and need to be executed in a console. The unique features of the comprehensive modules and nice user interface made MEGA3 a booming success. The paper on MEGA3 is highly cited (>800 in 2006) so that it increased the impact factor of the journal Briefings in Bioinformatics from around 4 to 24 (Bishop and Bird, 2007). The development of the evolutionary trace (ET) method tells a similar story from the other side. Although it was published as early as 1996, the powerful principle that underlies ET was not completely explored in its implementations, including two recent new ones (Lichtarge et al., 1996; Mihalek et al., 2006; Morgan et al., 2006). Only the inventor's group and their collaborators seem to make use of ET, although biologists consider ET as a well established method.

This clearly demonstrates that a user-friendly implementation may be even more appreciated by the user community than the original algorithm design. In my future research, I will continue to identify unmet research interests with my biological background. Then I will seek solutions first in computer science and design new algorithms if similar research has not been

done in computer science. Finally I aim to carefully implement the algorithms and provide a user-friendly graphic interface.

References

Bishop, M. and Bird, C. (2007) BIB's first impact factor is 24.37, *Brief Bioinform*, **8**, 207.

Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol*, **257**, 342-358.

Mihalek, I., Res, I. and Lichtarge, O. (2006) Evolutionary trace report_maker: A new type of service for comparative analysis of proteins, *Bioinformatics*, **22**, 1656-1657.

Morgan, D.H., Kristensen, D.M., Mittelman, D. and Lichtarge, O. (2006) ET viewer: An application for predicting and visualizing functional sites in protein structures, *Bioinformatics*, **22**, 2049-2050.

Summary

Each protein is characterized by its unique sequential order of amino acids, the so-called protein sequence. Biology”s paradigm is that this order of amino acids determines the protein”s architecture and function. In this thesis, we introduce novel algorithms to analyze protein sequences. Chapter 1 begins with the introduction of amino acids, proteins and protein families. Then fundamental techniques from computer science related to the thesis are briefly described. Making a multiple sequence alignment (MSA) and constructing a phylogenetic tree are traditional means of sequence analysis. Information entropy, feature selection and sequential pattern mining provide alternative ways to analyze protein sequences and they are all from computer science.

In Chapter 2, information entropy was used to measure the conservation on a given position of the alignment. From an alignment which is grouped into subfamilies, two types of information entropy values are calculated for each position in the MSA. One is the average entropy for a given position among the subfamilies, the other is the entropy for the same position in the entire multiple sequence alignment. This so-called two-entropies analysis or TEA in short, yields a scatter-plot in which all positions are represented with their two entropy values as x- and y-coordinates. The different locations of the positions (or dots) in the scatter-plot are indicative of various conservation patterns and may suggest different biological functions. The globally conserved positions show up at the lower left corner of the graph, which suggests that these positions may be essential for the folding or for the main functions of the protein superfamily. In contrast the positions neither conserved between subfamilies nor conserved in each individual subfamily appear at the upper right corner. The positions conserved within each subfamily but divergent among subfamilies are in the upper left corner. They may participate in biological functions that divide subfamilies, such as recognition of an endogenous ligand in G protein-coupled receptors.

The TEA method requires a definition of protein subfamilies as an input. However such definition is a challenging problem by itself, particularly because this definition is crucial for the following prediction of specificity positions. In Chapter 3, we automated the TEA method described in Chapter 2 by tracing the evolutionary pressure from the root to the branches of the phylogenetic tree. At each level of the tree, a TEA plot is produced to capture the signal of the evolutionary pressure. A consensus TEA-O plot is composed from the whole series of plots to provide a condensed representation. Positions related to functions that evolved early (conserved) or later (specificity) are close to the lower left or upper left corner of the TEA-O plot, respectively. This novel approach allows an unbiased, user-independent, analysis of residue relevance in a protein family. We tested the TEA-O method on a synthetic dataset as well as on “real” data, i.e., LacI and GPCR datasets. The ROC plots for the real data showed

that TEA-O works perfectly well on all datasets and much better than other considered methods such as evolutionary trace, SDPpred and TreeDet.

While positions were treated independently from each other in Chapter 2 and 3 in predicting specificity positions, in Chapter 4 multi-RELIEF considers both sequence similarity and distance in 3D structure in the specificity scoring function. The multi-RELIEF method was developed based on RELIEF, a state-of-the-art Machine-Learning technique for feature weighting. It estimates the expected "local" functional specificity of residues from an alignment divided in multiple classes. Optionally, 3D structure information is exploited by increasing the weight of residues that have high-weight neighbors. Using ROC curves over a large body of experimental reference data, we showed that multi-RELIEF identifies specificity residues for the seven test sets used. In addition, incorporating structural information improved the prediction for specificity of interaction with small molecules. Comparison of multi-RELIEF with four other state-of-the-art algorithms indicates its robustness and best overall performance.

In Chapter 2, 3 and 4, we heavily relied on multiple sequence alignment to identify conserved and specificity positions. As mentioned before, the construction of such alignment is not self-evident. Following the principle of sequential pattern mining, in Chapter 5, we proposed a new algorithm that directly identifies frequent biologically meaningful patterns from unaligned sequences. Six algorithms were designed and implemented to mine three different pattern types from either one or two datasets using a pattern growth approach. We compared our approach to PRATT2 and TEIRESIAS in efficiency, completeness and the diversity of pattern types. Compared to PRATT2, our approach is faster, capable of processing large datasets and able to identify the so-called type III patterns. Our approach is comparable to TEIRESIAS in the discovery of the so-called type I patterns but has additional functionality such as mining the so-called type II and type III patterns and finding discriminating patterns between two datasets.

From Chapter 2 to 5, we aimed to identify functional residues from either aligned or unaligned protein sequences. In Chapter 6, we introduce an alignment-independent procedure to cluster protein sequences, which may be used to predict protein function. Traditionally phylogeny reconstruction is usually based on multiple sequence alignment. The procedure can be computationally intensive and often requires manual adjustment, which may be particularly difficult for a set of deviating sequences. In cheminformatics, constructing a similarity tree of ligands is usually alignment free.

Feature spaces are routine means to convert compounds into binary fingerprints. Then distances among compounds can be obtained and similarity trees are constructed via clustering techniques. We explored building feature spaces for phylogeny reconstruction either using the so-called *k*-mer method or via sequential pattern mining with additional filtering and combining operations. Satisfying trees were built from both approaches

compared with alignment-based methods. We found that when *k* equals 3, the phylogenetic tree built from the *k*-mer fingerprints is as good as one of the alignment-based methods, in which PAM and Neighborhood joining are used for computing distance and constructing a tree, respectively (NJ-PAM). As for the sequential pattern mining approach, the quality of the phylogenetic tree is better than one of the alignment-based method (NJ-PAM), if we set the support value to 10% and used maximum patterns only as descriptors.

Finally in Chapter 7, general conclusions about the research described in this thesis are drawn. They are supplemented with an outlook on further research lines. We are convinced that the described algorithms can be useful in, e.g., genomic analyses, and provide further ideas for novel algorithms in this respect.

Samenvatting

Elk eiwit wordt gekenmerkt door zijn unieke opeenvolgende orde van aminozuren, de zogenaamde eiwitsequentie. Het paradigma van de biologie is dat deze volgorde van aminozuren de structuur en functie van het eiwit bepaalt. In dit proefschrift worden nieuwe algoritmen geïntroduceerd om eiwitsequenties te analyseren. Hoofdstuk 1 begint met een inleiding over aminozuren, eiwitten en eiwitfamilies. Daarna worden kort fundamentele technieken uit de informatica die verband houden met dit proefschrift besproken. Traditionele manieren om een eiwitsequentie te analyseren zijn het maken van een zogenaamde “multiple sequence alignment (MSA)” en het construeren van een fylogenetische boom. Informatie-entropie, eigenschap-selectie en het vinden van opeenvolgende patronen, allemaal technieken afkomstig uit de informatica, bieden andere manieren om eiwitsequenties te analyseren.

In Hoofdstuk 2 werd informatie-entropie gebruikt om de mate van conservering van een zekere positie in de “alignment” vast te stellen. Twee soorten informatie-entropie werden berekend voor elke positie in een MSA die zelf al onderverdeeld is in subfamilies. Eén is de gemiddelde entropie voor een positie in de subfamilies, de andere is de entropie voor dezelfde positie in de gehele MSA. Deze zogenaamde twee-entropieën-analyse (TEA) geeft een puntengrafiek waarin alle posities worden weergegeven door de twee entropiewaarden als x- en y-coördinaten. De verschillende posities van de punten in de grafiek staan voor de verschillende conserveringspatronen, en vormen even zovele suggesties voor verschillende biologische functies. De posities die globaal geconserveerd zijn verschijnen in de linkerbenedenhoek van de grafiek, hetgeen suggereert dat deze posities van het grootste belang zijn voor de vouwing of het algemeen functioneren van de eiwitsuperfamilie. In tegenstelling hiermee verschijnen in de rechterbovenhoek die posities die noch geconserveerd zijn tussen subfamilies noch binnen een subfamilie. De posities die geconserveerd zijn in een subfamilie maar verschillen tussen subfamilies verschijnen in de linkerbovenhoek. Deze posities kunnen een biologische functie hebben die verschilt tussen subfamilies, zoals de herkenning van een (endogeen) ligand in aan G eiwitten gekoppelde receptoren (GPCRs).

De definitie vooraf van eiwitsubfamilies is een vereiste als invoer voor de TEA methode. Zo”n definitie is op zichzelf echter al een stevig probleem, vooral omdat deze definitie cruciaal is voor de daaropvolgende voorspelling van “specificity” posities. In Hoofdstuk 3 werd daarom de TEA methode geautomatiseerd door de evolutionaire druk van de wortel tot aan de takken van de fylogenetische boom na te trekken. Op elk niveau van de boom wordt een TEA-grafiek gemaakt om het signaal van de evolutionaire druk te vangen. Van een hele reeks van dergelijke grafieken wordt dan een consensus TEA-O grafiek gemaakt als een gecondenseerde weergave van alle grafieken. Posities die te maken hebben met eiwitfuncties die eerder (“conserved”) of later (“specificity”) ontstonden, bevinden zich respectievelijk in de linker beneden- dan wel linker bovenhoek van de TEA-O grafiek. Deze nieuwe benadering

maakt een analyse van de betekenis van bepaalde residuen in een eiwitfamilie mogelijk die noch vooringenomen noch afhankelijk van de gebruiker is. We probeerden de TEA-O methode uit op een kunstmatige dataset en op “echte” gegevens, namelijk. datasets voor LacI-eiwitten en GPCRs. Met behulp van ROC grafieken voor deze laatste gegevens werd aangetoond dat TEA-O uitstekend werkt op alle datasets, en zelfs veel beter dan andere bestaande methoden zoals “evolutionary trace”, SDPpred en TreeDet.

Terwijl in de Hoofdstukken 2 en 3 de posities in de eiwitsequenties als onafhankelijk van elkaar werden beschouwd, wordt in hoofdstuk 4 multi-RELIEF besproken waarin zowel de gelijksoortigheid in sequentie als de afstand in de driedimensionale structuur wordt meegenomen. Deze nieuwe methode werd ontwikkeld op basis van RELIEF, een innovatieve “Machine Learning” techniek voor de weging van eigenschappen. Er wordt een schatting gemaakt van de verwachte “lokale” functionele specificiteit van de residuen gebaseerd op een “alignment” die in meerdere klassen onderverdeeld is. Als een optie worden gegevens van de driedimensionale structuur gebruikt door het gewicht van de residuen die zelf ook “zwaargewicht” buren hebben te verhogen. Met behulp van ROC grafieken van een grote hoeveelheid experimentele referentiegegevens konden we laten zien dat multi-RELIEF inderdaad “specificity” residuen identificeert voor alle zeven gebruikte test sets. Bovendien verbeterde het inbrengen van structurele informatie de voorspellingen met betrekking tot de specificiteit voor de interactie met kleine moleculen. Een vergelijking van multi-RELIEF met vier andere recente algoritmen wijst uit dat onze methode robuust is en over de gehele linie het best presteert.

In de Hoofstukken 2, 3 en 4 werd zwaar geleund op een “multiple sequence alignment” om de geconserveerde en “specificity” posities aan te tonen. Zoals al eerder vermeld is de constructie van zo”n “alignment” geen vanzelfsprekende zaak. Daarom presenteren we in Hoofdstuk 5 een nieuw algoritme dat, gebaseerd op de principes van het vinden van opeenvolgende patronen, direct, dus zonder eerst de sequenties te “alignen”, veel voorkomende en biologisch belangrijke patronen identificeert. Zes algoritmen werden ontworpen en geïmplementeerd om drie verschillende patroontypen in één of twee datasets te herkennen door middel van een benadering waarin naar steeds langer wordende patronen (“pattern growth”) wordt gezocht. We vergeleken onze benadering met PRATT2 en TEIRESIAS door te kijken naar efficiëntie, volledigheid en de diversiteit van patroontypen. Vergeleken met PRATT2 is onze methode sneller, in staat om grote datasets te verwerken, en om zogenaamde type III patronen aan te tonen. Onze methode is vergelijkbaar met TEIRESIAS als het gaat om de het ontdekken van de zogenaamde type I patronen, maar heeft extra functionaliteit als het gaat om het zoeken naar zogenaamde type II en III patronen, en het vinden van onderscheidende patronen tussen de twee datasets.

In de Hoofdstukken 2 tot en met 5 beoogden we functionele residuen te identificeren in eiwitsequenties die al of niet “gealigned” waren. In Hoofdstuk 6 introduceren we een

“alignment”-onafhankelijke procedure voor het clusteren van eiwitsequenties, die gebruikt kan worden voor het voorspellen van de functie van een eiwit. De reconstructie van fylogenie is traditioneel meestal gebaseerd op een “multiple sequence alignment”. Deze procedure kan echter veel computertijd kosten en moet nog vaak handmatig aangepast worden, hetgeen bepaald lastig kan zijn voor een reeks afwijkende sequenties. In de cheminformatica vindt de constructie van een gelijksoortigheidsboom van liganden over het algemeen plaats zonder “alignment”. Eigenschapruimten vormen een standaardmanier om verbindingen om te zetten in binaire vingerafdrukken. Daarna is het mogelijk om afstanden tussen verbindingen te berekenen en gelijksoortigheidsbomen te construeren door middel van clusteringtechnieken. We onderzochten twee methoden voor het verkrijgen van zulke eigenschapruimten ten behoeve van fylogeniereconstructie, de zogenaamde *k*-mer methode of één met daarin het vinden van opeenvolgende patronen met aanvullende filter- en combinatiebewerkingen. Met deze twee methoden konden bevredigende bomen worden gebouwd vergeleken met de “alignment” methodieken. Als *k* gelijk is aan 3, zo leerden we, dan is de fylogenetische boom gebouwd met de k-mer vingerafdrukken net zo goed als één van de op “alignment” gebaseerde technieken, waarin PAM en “Neighborhood joining” beide worden gebruikt om de afstanden tussen de sequenties te berekenen en de boom te construeren (NJ-PAM). Net zoals voor de benadering waarin opeenvolgende patronen worden gevonden is de kwaliteit van de fylogenetische boom beter dan de NJ-PAM methode, als de “support value” op 10% werd gezet en de maximumpatronen werden gebruikt als descriptoren.

Tot slot worden in Hoofdstuk 7 algemene conclusies over het onderzoek in dit proefschrift getrokken, die worden aangevuld met een vooruitblik naar verdere onderzoekslijnen. We zijn ervan overtuigd dat de beschreven algoritmen nog veel meer toepassingsmogelijkheden hebben, bijvoorbeeld in de analyse van het genoom, en geven ook ideeën voor nieuwe algoritmen hiervoor.

Publications grouped by fields

Bioinformatics: novel algorithms and implementations

- **Kai Ye**, Gert Vriend, Adriaan P. IJzerman.
  Tracing evolutionary pressure.
  ***Bioinformatics**. 2008, 24(7):908-915.*
- **Kai Ye**, K. Anton Feenstra, Jaap Heringa, Adriaan P. IJzerman, Elena Marchiori.
  Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine Learning approach for feature weighting.
  ***Bioinformatics**. 2008, 24(1):18-25.*
- **Kai Ye**, Walter A. Kosters, Adriaan P. IJzerman.
  An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences.
  ***Bioinformatics**. 2007, 23(6):687-693.*
- **Kai Ye**, Eric-Wubbo M. Lameijer, Margot W. Beukers, Adriaan P. IJzerman.
  A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors
  ***Proteins**: Structure, Function, and Bioinformatics. 2006, 63:1018-1030*
- **Kai Ye**, Walter A. Kosters, Adriaan P. IJzerman.
  Alignment independent phylogeny reconstruction – A cheminformatics approach.
  Submitted

Pharmacology modeling: mathematical modeling of receptor-ligand interaction

- Laura Heitman, Jacobus P. D. van Veldhoven, Annelien M. Zweemer, **Kai Ye**, Johannes Brussee, Ad P. IJzerman
  False positives in a reporter gene assay: identification and synthesis of substituted N-pyridin-2-ylbenzamides as competitive inhibitors of firefly luciferase.
  ***Journal of Medicinal Chemistry**. 2008, 51(15):4724-4729*
- Laura Heitman, **Kai Ye**, Julia Oosterom, Ad P. IJzerman
  Amiloride derivatives and a non-peptidic antagonist bind at two distinct allosteric sites in the human gonadotropin-releasing hormone receptor
  ***Molecular Pharmacology**. 2008, 73(6):1808-1815*
- Qilan Li, **Kai Ye**, Clara C. Blad, Hans den Dulk, Jaap Brouwer, Ad P. IJzerman, Margot W. Beukers.
  ZM241385, DPCPX and MRS1706 are inverse agonists with different relative intrinsic efficacies on constitutively active mutants of the human adenosine A_{2B} receptor.
  ***Journal of Pharmacology and Experimental Therapeutics**. 2007, 320(2):637-45.*
- Aniko Goblyos, Zhan-Guo Gao, Johannes Brussee, Roberto Connestari, Sabrina Neves Santiago, **Kai Ye**, Adriaan P. IJzerman, Kenneth A. Jacobson.
  Structure-activity relationships of new 1H-imidazo[4,5-c]quinolin-4-amine derivatives as allosteric enhancers of the A3 adenosine receptor.
  ***Journal of Medicinal Chemistry**. 2006 49(11):3354-3361*
- **Kai Ye**, Thea Mulder-Krieger, Margot W. Beukers, Ad P. IJzerman. [^{3}H]adenine's high filter binding precludes its use as a radioligand to study the adenine receptor.
  ***Purinergic Signaling.** 2006 2:71-72*

Molecular Biology

- Fethia Ben Yebdri, Abderrahmane AAZAZ, **Kai Ye**, Huiwen Ma, Liheng Tong
  Humoral immune response elicited by plasmid DNA containing HGV E2 gene fragment.

Chinese Journal of Biotechnology *2004 20(5):683-688.*

- Hong Xu, Xintian Lai, **Kai Ye**, Huiwen Ma, Kui Hong
  Expression of a DNA fragment encoding the active domain of human TNF related apoptosis inducing ligand in *pichia pastoris*.
  Chinese Journal of Biotechnology *2003 19(2):163-167.*
- **Kai Ye**, Qilan Li, Hong Xu, Kironde Fred Alexander S, Huiwen Ma
  Immune characteristics of plasmid DNA containing *Plasmodium falciparum* Pf70 gene fragment.
  Chinese Journal of Microbiology and Immunology *2003 23 (5):359-363*
- Zhuohua Wang, **Kai Ye**, Hong Xu, Huiwen Ma, Lihong Tong, Xiliang Peng
  Expression and characterization of envelope protein 2 gene of hepatitis G virus in *Pichia pastoris*.
  Chinese Journal of Biotechnology *2002 18(2):187-192*

Curriculum vitae

Kai Ye was born on the 19th of July 1977 in HuBei, China. In 1995, he started his university education at Wuhan University, China, majoring in biopharmaceutical science. In 1999, he continued his study as a master student at the College of Pharmacy at Wuhan University. During this stage, he performed an internship in the group of prof. Dr. Huiwen Ma, investigating a DNA vaccine for malaria. After graduation, he was immediately appointed as a lecturer in the College of Pharmacy, Wuhan University, teaching bioinformatics to master students. In the beginning of 2004, he attended a cooperation project between Chinese Universities and Leiden University and moved from China to the Netherlands. By the end of 2004, he obtained an MPhil (master of philosophy) degree at Leiden University. Starting in 2005, he continued his study in the lab of prof. Dr. Ad IJzerman (Medicinal Chemistry, LACDR). His PhD project was mainly focused on novel algorithms to analyze and extract useful information from protein sequences, particularly G protein-coupled receptors. Since 1st January 2008, he was appointed as a postdoctoral fellow in the IJzerman group. As of July 2008, he continued his postdoctoral study at the European Bioinformatics Institute in the United Kingdom to compute of unique substrings in human genome.

Acknowledgement

I owe a debt of gratitude to many people who helped me and supported me in the last 4 years. It is the right time to acknowledge all of them.
Everything that has an end must also have a starting point. The wisdom and charming character of Prof. D.D. Breimer formed my first impressions of both Leiden University and the Netherlands. I am deeply grateful to him for his support to pursue MPhil and PhD degrees at Leiden University.
My foremost thanks go to my supervisor prof. Ad IJzerman, for providing me such a great opportunity to pursue my PhD education in the division of Medicinal Chemistry. His inspiring encouragement, personal guidance and his valuable feedback contributed greatly to this thesis. Most importantly, he always encouraged me to explore new fields: pharmacology and bioinformatics. I really appreciate and value the confidence and trust he had in me which not only brought me interesting projects to complete this thesis but also will have a positive effect on my future career.
Secondly, I should like to thank my co-promoter Water Kosters. Without his guidance and patience, I would be completely lost in computer science and require 4,000 years to finish my PhD thesis simply because he showed me how to speed up my computer program 1,000 times.
Thirdly, I owe debt to Eric-Wubbo Lameijer. He helped me in writing my first useful and of course functional computer program. He made me believe that, even as a biologist, I am able to develop my skills in computer science.
I should acknowledge Thea for her great practical contribution in adenine project although this work is not present in this thesis. And I appreciate help of Rob who helped me to correct the English in part of this thesis.
I have to admit, life in a completely new place is not easy. Thanks to my friends and colleagues, the help I received made the life in these 4 years easier and smoother. Ad, Margot, Hans, Henk, Lisa, Leon, Jerone, Jacobus, Elke, Jurre, Marloes, Tao Yue, Xingfu Xu, Zhi Ding, Linghua Jiang, Haiyan Liu and Junjun Shan, I'll never forget your help.
Last but not least, I should like to thank my family for the love and support in all my pursuit. I feel deeply indebted to my parents that I did not take care of them since I was so far away from them. I very much appreciate the understanding, encouragement, support and cooperation of my wife Qilan Li.