In silico discoveries for biomedical sciences

Haagen, H. van

Chapter 5

Finding gene-disease relations using implicit information in the scientific literature

Herman van Haagen[1], Emmelien Aten[1], Peter-Bram 't Hoen[1], Marco Roos[1,2], Tobias Messemaker[?], Erik A. Schultes[1], Barend Mons[1], Gert-Jan van Ommen[1], and Martijn Schuemie[3]

1. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

2. Institute for Informatics, University of Amsterdam, Amsterdam, The Netherlands

3. Biosemantics Group, Erasmus Medical Center, Rotterdam, TheNetherlands

Manuscript in preparation

Abstract
Despite large and ever-growing bioinformatic data sets, there is often no information that explicitly links genes to a disease in literature. Bioinformatic approaches have attempted to circumvent this problem by searching for genes similar to those already known to be associated with a disease [1-4]. However, this approach is frequently not useful because previous associated genes with a disease are not available. Here, we use concept profiles [5, 6], a vector-based description of terms, to discover implied relationships between genes and diseases for which no explicit link (co-occurrence) has been stated in either text or any other database. In a retrospective text mining analysis of scientific literature concept profiles were able to prioritize disease genes on average within the top 13 out of 200 genes located in a specified linkage interval at least one year before the publication of the landmark paper explicitly establishing the gene-disease relationship. Examination of the highly-ranked concepts shared between the gene and the disease in concept profiles was used by biomedical experts to evaluate the plausibility of the inferred relationships and rationalize potential biological mechanisms. By exploiting the implicit information in the literature, concept profiles performed two-fold better in prioritizing genes of polygenic diseases than the Endeavour gene prioritizer [2] using 26 data-mining resources. These results demonstrate the enormous untapped potential of implied information in scientific literature for biomedical discovery, and the application of concept profile technology in extracting new knowledge.

Introduction

Although linkage analysis, association studies, and next generation sequencing technology have produced voluminous amounts of genetic data that are essential for the characterization of disease mechanisms, isolating genes that cause or impact the etiology of a particular disease remains a time consuming and largely serendipitous task. Often, many interrelated factors must be considered. For example, individual genes may cause multiple diseases, distinct diseases may be caused by multiple genes, and different diseases will often have phenotypic overlap. To cope with these inherent complexities and with the size of large and rapidly growing datasets, bioinformatic tools have been developed combining text-mining and data-mining capabilities to automatically search for correlations among, and then prioritize, putative gene-disease pairs[7-14]. For example, the Endeavour web tool combines biomedical ontologies, text, and data from 26 distinct sources to prioritize genes for specific diseases [2]. Many of the prioritizers that integrate multiple data sources are based on so called ‘seed’ genes, which are genes having a known relation to a disease that help to find the next causative gene that results in the same phenotype. For instance, a novel gene for breast cancer may be found by using information about BRCA1 and BRCA2, genes already known to

cause the disease. However, for the majority of diseases recorded in OMIM, a causative gene is not yet known. In these cases, prioritizers based on seed genes do not work. Furthermore, for all those diseases or syndromes where the first gene has yet to be discovered, a prioritizer will be limited to text information only, yet before a landmark paper is published describing the disease causing gene, the disease and the gene tend to have few or no co-occurrence in the same abstract or article. Therefore text-mining systems based on direct co-occurrences will fail to predict the majority gene-disease relationships.

However, woven within the narrative of scientific literature there are a vast network of relations among terms that are to some degree left implicit by the authors. Implicit relations may arise as a consequence of new findings or as part of the scientific rational, and may or may not be intentional. Implicit information may be directly related to the immediate narrative or may have ancillary relations. Here, we used a text-mining method based on concept profiles to prioritize candidate genes by considering this large amount of implicit associative information in text. A concept profile for a given concept contains all other concepts that have a co-occurrence weighted by the Uncertainty Coefficient [5]. Concept profiles must be constructed uniquely for a given ontology and corpus [15, 16], but once they have been constructed, the similarity between any two concept profiles can be computed by taking the inner product of their corresponding weights, the so-called match score[17]. The statistical significance of the match score between the profiles of two concepts (i.e., gene and disease) can be evaluated by comparing the log transform of the match scores to that of a null distribution constructed from randomly chosen concept pairs (Fig 1). Hence, it is possible using concept profiles to establish a statistically significant association between concepts based on highly ranked concepts in their profiles, even when they do not have a co-occurrence (*i.e.* usually an explicit stated relationship) in the literature. Discovery of novel and informative associations between genes and diseases is thus not dependent on linkage analyses or seed genes.

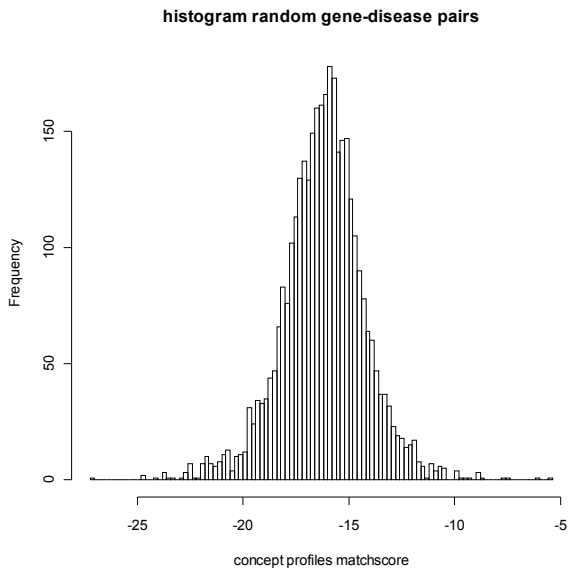

Figure 1. Distribution of concept profiles match scores calculated for randomly chosen gene-disease combinations. The MEDLINE abstract text corpus is from 1980 until May 2009.

Results and Discussion

We evaluated the effectiveness of concept profiles using 18 previously described gene-disease relationships taken from the Human Reference Protein Data base (HRPD)[18] (Table 1). Concept profiles for the genes and the diseases were constructed from all MEDLINE abstracts up to one year before the landmark paper explicitly describing the link between the gene and disease was published. This roll-back analysis used two time-delimited corpora: From 1980 to February 2005 (for landmark publications dating from February 2006 to December 2006) and from 1980 to August 2006 (landmark publications appearing after august 2007). For each of these gene-disease pairs, no co-occurrence was found between the gene and the disease before the landmark paper was published, both the gene and the disease appear in a minimum of five abstracts and the disease is currently considered to be monogenic. For each test gene an artificial linkage interval was arbitrarily set containing 200 genes (100 genes upstream and downstream of the test gene) following the approach by Aerts et. al. [7].

Table 1. Gene-disease pairs using concept profiles.

Gene	Disease	Landmark Publication Date	PMID	Rank	p-value
MFN2	Hereditary motor and sensory neuropathy VI	February 2006	16437557	2	0.0018
RECQL4	Baller-Gerold syndrome	February 2006	15964893	1	0.0046
KRT85	Ectodermal dysplasia, pure hair-nail type	March 2006	16525032	13	0.0093
ACVR1	Fibrodysplasia ossificans progressiva	May 2006	16642017	2	0.0052
TGFBI	Corneal dystrophy, epithelial basement membrane	June 2006	16652336	1	0.00045
IL10RB	Hepatitis B virus, susceptibility to	June 2006	16757563	16	0.062
IFN-AR2	Hepatitis B virus, susceptibility to	June 2006	16757563	3	0.0065
PLA2G6	Infantile neuroaxonal dystrophy 1	July 2006	16783378	11	0.043
TREX1	Aicardi-Goutieres syndrome 1	August 2006	16845398	105	0.93
CHRNG	Escobar syndrome	August 2007	16826520	66	0.77
DOK7	Myasthenia, limb-girdle, familial		16917026	NaN	NaN
SCN9A	Paroxysmal extreme pain disorder	September 2006	17145499	15	0.084
MYH11	FAA4	March 2006	16444274	4	0.013
TFAP2A	branchio-oculo-facial syndrome	May 2008	18423521	4	0.068
PIK3CA	Seborrheic keratosis	August 2007	17673550	9	0.076
VLDLR	Dysequilibrium syndrome	February 2008	18043714	27	0.18
BUB1B	PCS	February 2006	16411201	12	0.29
TRPV4	Brachyolmia	August 2008	18587396	44	0.76
			Average rank 20		Average p

The concept profile for each gene in the linkage area was matched with the disease profile resulting in a ranked list of the test gene among the 200 genes in the artificial linkage interval (Table 1). On average the test genes ranked within the top 20, and in two cases (epithelial basement membrane corneal dystrophy (EBMD)) and Baller-Gerold syndrome), the test genes ranked number one. However, the TGFBI gene is often co-mentioned with generic disease types, like hereditary corneal dystrophy and corneal dystrophy (column 3 in Table 2) suggesting that its high rank is not necessarily an indicator of a specific relation to EBMD. For the 'Myasthenia, limb-girdle, familial' there was not enough information for the test gene to build a concept profile. When prioritizing gene-disease pairs in practice, it is essential that the significance of the putative gene-disease relation be subject to evaluation. Hence one-sided p-values were calculated from the concept profile

match scores, using the null distribution (Fig 1 & Table 1). This p-value is indicative of the quality of the prediction, and therefore of the reliability of the ranking of the list: the cases with very high ranks could have been predicted based on the low p-value of the gene (Fig 2). If we had used a cut-off p-value of 0.02 to reject the prioritizer output the results for Aicardi-Goutieres syndrome 1, Brachyolmia and Myasthenia, limb-girdle, familial would have been rejected. Escobar syndrome (test gene ranks 66) with a p-value of 0.017 for the highest ranked gene would have remained as a reliable output. With three outliers rejected the average rank of the remaining 15 samples would becomes 12.5. Clearly concept profiles are highly effective in identifying gene-disease pairs deliberately using only the implicit information in MEDLINE prior to the landmark paper.

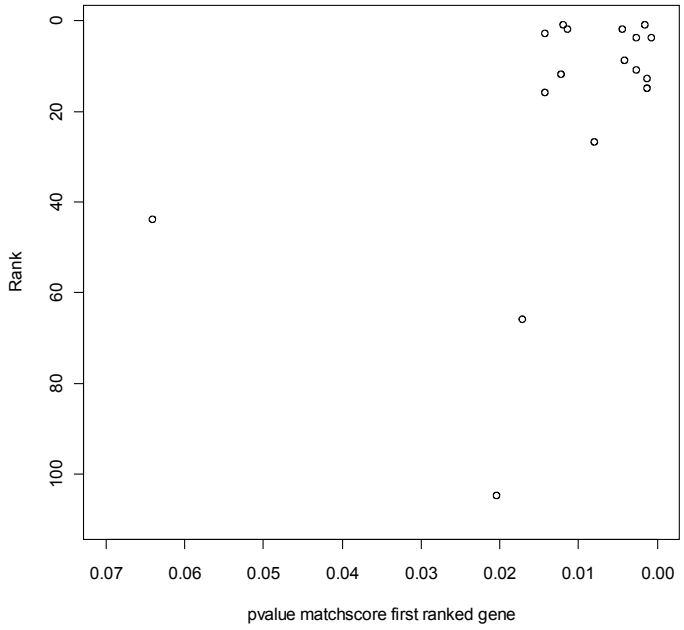

Figure 2. p-value of the highest rank gene versus the rank of the test gene.

In addition to prioritizing genes, concept profiles provide important biological insight revealing how the gene might be associated with a disease. However, by inspecting the highly ranked concepts in the two concept profiles that linking the gene and disease a biomedical expert would likely (example of gene *PIK3CA*

Table 1) choose for instance gene with rank nine over of the first eight for further investigation. To explore the utility of the information in concept profiles in rationalizing predicted gene-disease pairs we chose the three gene-disease pairs having the highest ranking: Hereditary motor and sensory neuropathy VI (HMSN VI), Baller-Gerold syndrome (B-G syndrome), and EBMD. Biomedical researchers with expertise in these genes and diseases evaluated the shared concepts in concept profiles for their biological significance. Table 2 shows the top five of overlapping concepts between the gene and disease concept profiles. For B-G syndrome the dominating concept is Rothmund-Thomson (R-T) syndrome (contributes more than 95% to the overall score). Two documents were found that support the association between B-G syndrome and R-T syndrome, PMID: 11045594 and 9934984. The first one gives information for the clinical phenotypic overlap between the two syndromes. The gene RECQL4 has been co-mentioned before with R-T syndrome as a gene that when mutated causes this syndrome (PMID: 12379465, 12601557, 12673665, 12734318, 12838562, 12915449, 12952869, 15221963, 15317757, and 15558713). Because of the phenotypic overlap between the two syndromes, RECQL4 would be the most likely candidate to investigate first. Indeed, the landmark paper (PMID: 15964893) reports precisely this reasoning: "Clinical overlap between BGS, Rothmund-Thomson syndrome (RTS), and RAPADILINO syndrome is noticeable. Because patients with RAPADILINO syndrome and a subset of patients with RTS have RECQL4 mutations, we reassessed two previously reported BGS families and found causal mutations in RECQL4 in both."

Table 2. Indirect concepts linking the gene with the disease.

	Baller-Gerold syndrome		**Hereditary motor and sensory neuropathy VI**		**Corneal dystrophy, epithelial basement membrane**	
Top	Overlapping concepts	Contribution to score	Overlapping concepts	Contribution to score	Overlapping concepts	Contribution to score
1	rothmund-thomson syndrome	95.79	opa1	40.37	hereditary corneal dystrophy	42.28
2	Poikiloderma	2.47	optic atrophy, autosomal dominant	35.32	Corneal dystrophy	41.43
3	online mendelian inheritance in man	0.45	OPA1	23.17	lattice corneal dystrophy	12.08
4	Growth deficiency	0.32	Axonal neuropathy	0.61	Dystrophy	2.73
5	Clinical variability	0.24	recessive inheritance	0.3	Corneal erosion	0.58

In the case of HMSN VI, this disease is caused by mutations in the MFN2 gene. The overlapping concepts in the top three are all a form of optic atrophy 1. The first concept opa1 is the gene in zebrafish, the second concept is a disease and the third concept the human gene. Together they contribute more than 98% to the overall score. The landmark paper (PMID: 16437557) clearly shows that this concept is a strong indirect link, stating: "It is intriguing that MFN2 shows functional overlap with optic atrophy 1 (OPA1), the protein underlying the most common form of autosomal dominant optic atrophy, and mitochondrial encoded oxidative phosphorylation components as seen in Leber's hereditary optic atrophy." The MFN2 gene ranked second place (Table 1). This means one false positive before the test gene is found. The gene that ranks first place is KIF1B where the top five concepts between it and the disease are hereditary motor and sensory neuropathies (65.61%), HMSN II (15.76%), hereditary liability to pressure palsies (7.8%), Axonal neuropathy (4.61%), and HMSN I (2.96%). In consulting the supporting documents for KIF1B it was found that mutations Charcot-Marie-tooth disease type 2A1 (CMT 2A1 or HMSN2A1). Intriguingly, HMSN VI is also known as Charcot-Marie-tooth disease type 6 (CMT6). Thus, it appears that KIF1B is not a false positive but a gene that causes a related disease.

Detailed expert analyses of the concept profiles in linking genes to Seckel syndrome are provided in the Supplementary Information.

Table 3. Endeavor Gene Prioritizer predictions for monogenic and polygeneic diseases. Averages are only calculated over the ranks that are both covered by Endeavour and concept profiles.

Disease (monogenic)	**Endeavour**	**Concept profiles**
arrthythmia	4	20
congenital heart disease	(3)	NaN
cardiomyopathy 1	2	2
parkinsons disease	(50)	NaN
charcot-marie-tooth disease	14	1
amyotrophic lateral sclerosis	27	16
klippel-trenaunay disease	(3)	NaN
cardiomyopathy 2	1	10
distal hereditary motor neuropathy	15	51
Cornelia de Lange syndrome	(9)	NaN
average ranking	11	17
Disease (polygenic)	**Endeavour**	**Concept profiles**

Rheumatoid_arthritis	11	24
Parkinson_disease	23	30
Atherosclerosis1	54	5
Atherosclerosis2	29	21
Crohn disease	71	11
Alzheimer disease	54	3
Average ranking	40	16

To gauge the performance of concept profiles against methods based on co-occurrence, we replicated a recent study [2] using the gene prioritizer Endeavour where gene-disease predictions were made for ten monogenic and six polygenic diseases (Table 3). We generated concept profiles for the diseases in these test sets and for the test genes in their corresponding linkage interval. We used the same roll back analyses as Endeavour, taking only literature information up to one year before the landmark paper was published. For the monogenic diseases there were three genes where there was not enough information to calculate a match score using concept profiles. Of the 7 remaining gene-disease pairs for monogenic conditions, the average performance of the two methods was comparable. However, in the case of polygenic diseases having inherently complex interrelations among numerous genes and other concepts, concept profiles outperformed Endeavour's ranking on average by two-fold. By drawing on the deep network of conceptual relations that inform the study of polygenic diseases but usually remain un-stated in the literature, concept profiles are uniquely suited for knowledge discovery in complex multifactorial systems.

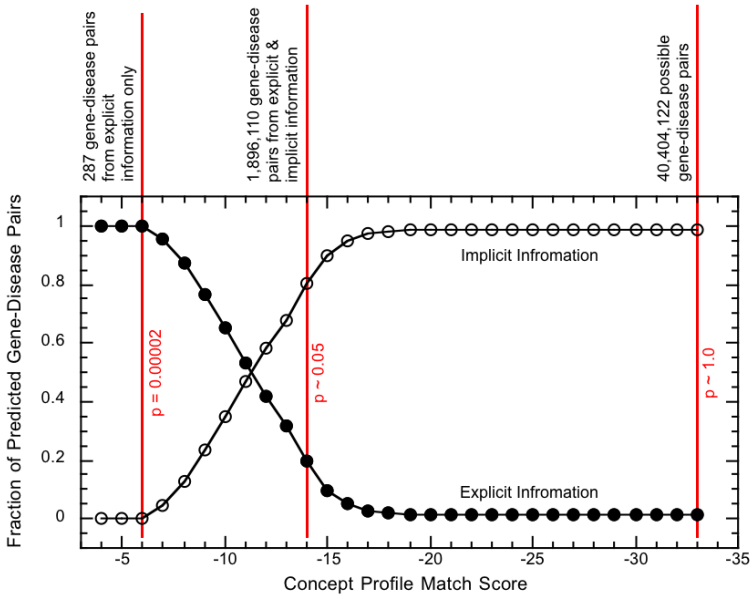

Figure 3. Estimation of implicit and explicit information. Co-occurrence methods can prioritize only 287 of the possible 40 million gene-disease pairs, while concept profiles can prioritize 5% at p= 0.05. Note the vast majority of textual information is implicit.

These results indicate the importance of implicit information in discovering new knowledge. Concept profiles can be used to estimate the relative proportions of implicit and explicit information. For example, given the number of genes and diseases meeting our minimal information criteria used herein, there are 40,404,412 possible gene-disease combinations. The match score and corresponding p-value for each these gene-disease pair can be calculated. For each p-value, the cumulative number of implicit and explicit gene-disease pairs (and then normalized to a percentage) can be computed (Supp Info Table 4). Thus, for each p-value, we know the fraction of the predicted pairs that are due to implicit information (Fig 3). For p=0.0 only gene-disease pairs with minimally one co-occurrence are found. But even for extremely significant p-values (p=0.00002) we already find some gene-disease pairs for which their association is only due to implicit information (*i.e.*, no co-occurrences found in MEDLINE). For p=0.003, still a highly significant gene-disease p value, the amount of implicit information is already 47%. For commonly accepted p-values around p=0.05, 88% of the gene-disease pairs are due to implicit information. Conclusion: The vast majority of

useful information in text is implicit, and this information is accessible with concept profiles.

There are 5330 gene-disease predictions that are better than p = 0.0002. To facilitate the expert evaluation of these predicted gene-disease pairs, the shared concepts from the concept profiles have been posted online along with the related PubMed IDs. Experts can search this data on gene or disease or any other related concept, and can provide their estimation of the quality of the prediction and leave commentary regarding possible biological mechanisms.

References

1. Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., et al., *Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.* Am J Hum Genet, 2006. **78**(6): p. 1011-25.
2. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
3. Oti, M., Snel, B., Huynen, M.A., and Brunner, H.G., *Predicting disease genes using protein-protein interactions.* J Med Genet, 2006. **43**(8): p. 691-8.
4. Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.
5. Jelier, R., Schuemie, M.J., Roes, P.J., van Mulligen, E.M., and Kors, J.A., *Literature-based concept profiles for gene annotation: the issue of weighting.* Int J Med Inform, 2008. **77**(5): p. 354-62.
6. Jelier, R., Schuemie, M.J., Veldhoven, A., Dorssers, L.C., Jenster, G., et al., *Anni 2.0: a multipurpose text-mining tool for the life sciences.* Genome Biol, 2008. **9**(6): p. R96.
7. Perez-Iratxeta, C., Bork, P., and Andrade, M.A., *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
8. van Driel, M.A., Cuelenaere, K., Kemmeren, P.P., Leunissen, J.A., and Brunner, H.G., *A new web-based data mining tool for the identification of candidate genes for human genetic disorders.* Eur J Hum Genet, 2003. **11**(1): p. 57-63.

9. Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B., et al., *Integration of text- and data-mining using ontologies successfully selects disease gene candidates.* Nucleic Acids Res, 2005. **33**(5): p. 1544-52.
10. Kohler, S., Bauer, S., Horn, D., and Robinson, P.N., *Walking the interactome for prioritization of candidate disease genes.* Am J Hum Genet, 2008. **82**(4): p. 949-58.
11. Radivojac, P., Peng, K., Clark, W.T., Peters, B.J., Mohan, A., et al., *An integrated approach to inferring gene-disease associations in humans.* Proteins, 2008. **72**(3): p. 1030-7.
12. Linghu, B., Snitkin, E.S., Hu, Z., Xia, Y., and Delisi, C., *Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network.* Genome Biol, 2009. **10**(9): p. R91.
13. Day, A., Dong, J., Funari, V.A., Harry, B., Strom, S.P., et al., *Disease gene characterization through large-scale co-expression analysis.* PLoS One, 2009. **4**(12): p. e8491.
14. Ala, U., Piro, R.M., Grassi, E., Damasco, C., Silengo, L., et al., *Prediction of human disease genes by human-mouse conserved coexpression analysis.* PLoS Comput Biol, 2008. **4**(3): p. e1000043.
15. Schuemie, M.J., Jelier, R., and Kors, J.A. *Peregrine: Lightweight gene name normalization by dictionary lookup*. in *Biocrative 2 workshop*. 2007. Madrid.
16. Tuason, O., Chen, L., Liu, H., Blake, J.A., and Friedman, C., *Biological nomenclatures: a source of lexical knowledge and ambiguity.* Pac Symp Biocomput, 2004: p. 238-49.
17. van Haagen, H.H.H.B.M., t Hoen, P.A.C., Botelho Bovo, A., de MorrÃ©e, A., van Mulligen, E.M., et al., *Novel Protein-Protein Interactions Inferred from Literature Context.* PLoS ONE, 2009. **4**(11): p. e7894.
18. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.

Supplementary information

Seckel syndrome is known as a rare autosomal recessive disorder characterized by growth retardation, microcephaly with mental retardation, and a characteristic 'bird-headed' facial appearance. Presently, only one gene in OMIM, ataxia-telangiectasia, is related to Seckel Syndrome via its mutated from RAD3-related protein (ATR)[1]. Recently a second gene that encodes for Centromere protein J (CENPJ) has been identified by Al-Dosari et. al.[2], but this gene had yet to be entered in OMIM at the time this analysis was completed. Of the top 20 proteins (out of the 12,391 proteins that had sufficient information to build a concept profile) having the highest match score to Seckel Syndrome, CENPJ ranks number 14, although CENPJ has no co-occurrences with Seckel syndrome in PubMed abstracts (Table 1). The concept microcephaly contributes the most to the match score and is the strongest implicit (or indirect) link between Seckel syndrome and CENPJ. Other candidate genes appear in this list that have been co-mentioned with Seckel syndrome before. For instance the protein pericentrin (PCNT, ranks 2) has three articles. The article with PMID: 18157127 describes that PCNT is another gene that causes Seckel syndrome. The article with PMID:19546241 gives a nice overview of related diseases with similar phenotype such as 'Primary microcephaly' and 'microcephalic osteodysplastic primordial dwarfism type II' (MOPD II). This article also lists microcephalin (MCPH1, ranks 5) as another disease-causing candidate. The last article (PMID: 16278902) also mentions the concept MOPD II. These results prompted us to further investigated whether CENPJ might be associated with PCNT and ATR. We generated a prioritized list for ATR and PCNT and checked the rank of CENPJ. Surprisingly CENPJ also showed no co-occurrences with ATR and it ranked 706. However, CENPJ is co-mentioned once with PCNT (PMID: 18174396). In this article PCNT is given as the cause for primordial dwarfism, and other candidate genes for Seckel syndrome are postulated (CDK5RAP2 ranks 32 not in table 1, and ASPM ranks 15). We performed a new search where related concepts for Seckel syndrome were used as PubMed input query and the results aggregated into a single rank using (Table 2). In this case, CENPJ ranks 4th and although ATR is a known gene for Seckel syndrome (recorded in OMIM), its information content is poor compared to the other related Seckel syndrome concepts.

From the PubMed abstracts we selected candidate genes that have been co-mentioned with Seckel syndrome related concepts and used them as a search query in the STRING database of known and predicted protein-protein interactions to see if there are any relations between these candidate genes (Fig 1). Again, ATM has many links with ATR, while all other links are mainly in the network of PCNT. CENPJ has a known physical interaction with PCNT. From a biological view, it would be highly interesting to identify the missing link between the ATR and the

PCNT pathway. In seeking potential relations between Seckel syndrome and CENPJ, it is clearly much easier to inspecting the concept profile overlap between (Table 3) than performing multiple PubMed search queries manually reading up to 17 articles. Lastly, this case demonstrates that using conventional co-occurrence approaches to predicting gene-disease relations could have negative performance results. Here, ATR, although the first choice to use as seed gene when looking for additional genes related to Seckel syndrome, would lead to false negative conclusions.

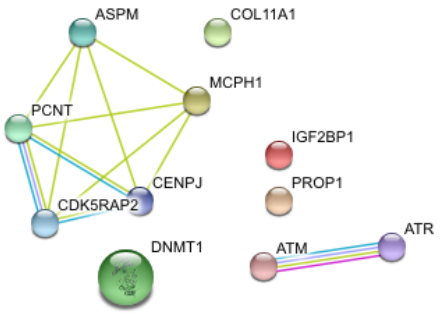

Figure 1. Candidate genes for Seckel syndrome in a network graph generated by STRING.

Supp Info Table 1. Prioritized list of proteins match with the profile of Seckel Syndrome. The column 'main concept' gives information which concept contributes the most to the score and is the strongest implicit (or indirect) link between Seckel syndrome and CENPJ. NUP85 is a homonym for PCNT and retrieves the same articles for PCNT.

rank	Co-occurrences	genename	Main concept	Contribution (%)	OMIM gene	PMIDs
1	7	ATR	ATR	74.43	x	[12640452, 14571270, 15309689, 15496423, 15616588, 16015581, 19504344]
2	3	PCNT	Seckel syndrome	54.38	x	[16278902, 18157127, 19546241]
3	2	NUP85	Seckel syndrome	74.54		[18157127, 19546241]
4	2	ANTXR1	ANTXR1	66.28		[12640452, 19504344]
5	3	MCPH1	MCPH1	61.84	x	[16217032, 17102619,

						19546241]
6	2	NBN	NBN	64.58	x	[15616588, 18664457]
7	0	MCPH2	Primary microcephaly	32.01		
8	2	FANCD2	FANCD2	47.9	x	[15314022, 15616588]
9	0	ATRIP	ATR	92.91		
10	1	DNMT1	DNMT1	98.49		[17015478]
11	1	MDC1	MDC1	19.13		[18664457]
12	1	FANCC	Fanconi's Anemia	31.02	x	[10232749]
13	6	PALB2	Fanconi's Anemia	36.35		[3115102, 6465473, 7686032, 10232749, 15314022, 17224058]
14	0	CENPJ	Microcephaly	17.77	x	
15	0	ASPM	Microcephaly	48.12	x	
16	5	CHEK1	CHEK1	28.75		[15616588, 16217032, 17015478, 18077418, 19504344]
17	0	PROP1	dwarfism	98.96	x	
18	2	MMAB	MMAB	60.9	x	[15314022, 19504344]
19	0	TOPBP1	ATR	64.53		
20	1	FOXL2	FOXL2	59.56	x	[16015581]

Table 2. Rank of proteins in prioritized lists for different concepts that are associated with Seckel syndrome, including Seckel syndrome itself

Gene name	Rank	Seckel syndrome	PCNT	MOPD II	Primary microcephaly	Microcephaly	ATR
PCNT	1	2	1	1	11	70	845
MCPH2	2	7	53	5	1	4	540
ASPM	3	15	38	6	2	2	622
CENPJ	4	14	18	4	3	5	706
ATR	5	1	385	53	29	49	1
MCPH1	6	5	64	7	4	12	271
NUP85	7	3	9	2	15	84	831
NBN	8	6	660	125	24	1	10
MDC1	9	11	233	15	6	63	15
TOPBP1	10	19	257	17	10	237	5
CDK5RAP2	11	32	36	10	5	14	1233
CHEK1	12	16	254	36	20	223	3
TP53BP1	13	27	500	26	13	170	16
RHO	14	41	373	21	8	261	28
CHEK2	15	26	428	87	36	177	4
ERCC2	16	23	1054	48	42	174	9
MRE11A	17	22	1118	584	63	9	11
GCP3	18	39	4	3	31	1077	7133

RAD50	19	25	881	478	67	10	17
SEH1L	20	104	54	12	89	67	753

Table 3. Overlapping concepts between Seckel syndrome and CENPJ

Top	Overlapping concept	Contribution (%)
1	Microcephaly	17.77
2	Primary microcephaly	17.31
3	MCPH1	11.86
4	Mcph1	11.44
5	MCPH1	11.44
6	MCPH1	11.41
7	MCPH1	7.54
8	PCNT	4.38
9	osteodysplastic primordial dwarfism	1.94
10	NUP85	1.11
11	MOPD II	0.93
12	pericentrin	0.77
13	dwarfism	0.72
14	Centrosome	0.55
15	Genes, Recessive	0.32

pvalue	matchscore	implicit	explicit	cum imp	cum exp	% imp	% exp
0	-4	0	29	0	29	0.00	1.00
0	-5	0	287	0	316	0.00	1.00
0.00002	-6	5	1139	5	1455	0.00	1.00
9.01E-05	-7	173	2341	178	3796	0.04	0.96
0.00023	-8	997	4333	1175	8129	0.13	0.87
0.00056	-9	3994	8561	5169	16690	0.24	0.76
0.00127	-10	12863	16612	18032	33302	0.35	0.65
0.00308	-11	38259	30653	56291	63955	0.47	0.53
0.00724	-12	109383	55429	165674	119384	0.58	0.42
0.01756	-13	327773	113012	493447	232396	0.68	0.32
0.04645	-14	1027450	142817	1520897	375213	0.80	0.20
0.11323	-15	2610325	78691	4131222	453904	0.90	0.10
0.23868	-16	5050536	26734	9181758	480638	0.95	0.05
0.41924	-17	7252607	6287	16434365	486925	0.97	0.03
0.60682	-18	7619189	906	24053554	487831	0.98	0.02
0.75026	-19	5790456	40	29844010	487871	0.98	0.02
0.84564	-20	3816141	0	33660151	487871	0.99	0.01
0.90624	-21	2460978	0	36121129	487871	0.99	0.01
0.94567	-22	1586840	0	37707969	487871	0.99	0.01
0.97008	-23	995043	0	38703012	487871	0.99	0.01
0.98434	-24	577399	0	39280411	487871	0.99	0.01
0.99236	-25	322237	0	39602648	487871	0.99	0.01
0.99673	-26	169158	0	39771806	487871	0.99	0.01
0.99855	-27	83531	0	39855337	487871	0.99	0.01
0.99948	-28	36376	0	39891713	487871	0.99	0.01
0.99972	-29	15168	0	39906881	487871	0.99	0.01
0.99982	-30	6123	0	39913004	487871	0.99	0.01
0.99989	-31	2528	0	39915532	487871	0.99	0.01
0.99993	-32	971	0	39916503	487871	0.99	0.01
0.99993	-33	38	0	39916541	487871	0.99	0.01

Table 4. Estimation of Implicit and Explicit Information

1. O'Driscoll, M., Ruiz-Perez, V.L., Woods, C.G., Jeggo, P.A., and Goodship, J.A., *A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome.* Nat Genet, 2003. **33**(4): p. 497-501.
2. Al-Dosari, M.S., Shaheen, R., Colak, D., and Alkuraya, F.S., *Novel CENPJ mutation causes Seckel syndrome.* J Med Genet. **47**(6): p. 411-4.