Text mining and computational chemistry reveal trends in applications and applicability of capillary electrophoresis

Palmblad, M.; Vleugels, R.; Bergquist, J.

Trends in Analytical Chemistry 159 (2023) 116946

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

journal homepage: www.elsevier.com/locate/trac

Text mining and computational chemistry reveal trends in applications and applicability of capillary electrophoresis

Magnus Palmblad [a, *], Reinier Vleugels [b], Jonas Bergquist [c]

[a] Center for Proteomics and Metabolomics, Leiden University Medical Center, PO Box 9600, 2300 RC, Leiden, the Netherlands
[b] Department of Computer Science, IBIVU Centre for Integrative Bioinformatics, Vrije Universiteit, Amsterdam, 1081 HV, the Netherlands
[c] Analytical Chemistry and Neurochemistry, Department of Chemistry - BMC, Box 599, Uppsala University, SE 75124, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 30 September 2022
Accepted 18 January 2023
Available online 21 January 2023

Keywords:
Capillary electrophoresis
Bibliometrics
Text mining
Machine learning
QSPR
Visualization

ABSTRACT

Capillary electrophoresis has matured into a highly sensitive and widely applied analytical method over the last forty years. Here we combine text mining and computational chemistry to paint, with very broad strokes, the applicability and trends in the scientific literature on capillary electrophoresis, simultaneously demonstrating that this is not only possible, but reveal both expected and unexpected details of this history. All software and data are freely available on GitHub (https://github.com/ReinV/SCOPE) and OSF (https://osf.io/e56zt/).

1. Introduction

Over the last four decades, capillary electrophoresis has grown into a sensitive and versatile technique. Variants of capillary electrophoresis have been applied to the analysis of small ions up to proteins, protein complexes and even as a separation tools for individual cells and organelles. In traditional reviews of a topic this large and diverse, spanning many thousands of publications, one would have to select the work deemed most relevant and hone in on a manageable (readable) number of papers to support or illustrate a particular narrative [1]. This process is inherently biased and relies on the expertise of the reviewers. While such bias is often valuable, recent advances in open access and artificial intelligence, when combined, enable knowledge extraction from the scientific literature on a large scale in a much less biased manner. In this Special Issue paper, we use this powerful combination to illuminate the history of the entire field of capillary electrophoresis, specifically looking at the applicability of different capillary electrophoresis methods to different classes of analytes over time.

2. Methods

To investigate the applicability of capillary electrophoresis techniques, we looked in existing text-mined annotations across 41.1 million titles and abstracts, and 8.2 million full-text papers, for chemicals co-occurring with capillary electrophoresis methods. We then connected the named entity recognitions (NERs) from the text mining with calculated or machine-learned predictions of physicochemical properties such as molecular weight, aqueous solubility and polarity (log *P*), as previously demonstrated [2], using the new Search and Chemical Ontology Plotting Environment (SCOPE) software pipeline [3].

The text-mined chemicals are those in the Chemical Entities of Biological Interest (ChEBI) ontology [4]. This ontology describes chemical compounds frequently reported in the scientific literature, including atoms, small molecules, ions and complexes, up to and including some well-known peptides. The analytical techniques, including specific capillary electrophoresis techniques and synonyms, were taken from the CHemical Methods Ontology (CHMO). This ontology describes chemical methods used to collect data, such as separation methods, spectroscopies and mass spectrometry. Specifically, we searched for “capillary electrophoresis” and separately each individual leaf node under this term in CHMO, namely “capillary affinity electrophoresis” (CAE), “capillary isoelectric focusing” (CIEF), “capillary isotachophoresis” (CITP) and

* Corresponding author.
E-mail address: n.m.palmblad@lumc.nl (M. Palmblad).

https://doi.org/10.1016/j.trac.2023.116946

"capillary gel electrophoresis" (CGE). Though not listed as capillary electrophoresis techniques in CHMO, we also looked for micellar electrokinetic capillary chromatography (MECC) and micellar electrokinetic chromatography (MEKC). Only the full name, within quotes for exact matches, were used, as the acronyms are ambiguous (e.g. "CGE" can also refer to "glassy carbon electrode"). We allowed matching anywhere in the papers, but also searched specifically in the methods sections in the subset of open access papers allowing tagged section searches.

With this structured list of literature search queries, we used SCOPE to access Europe PMC through its *articles* and *annotations* APIs [5] and automatically retrieve all text-mined chemical entities from those publications that match the search queries (out of the 41.1 million titles and abstracts and 8.2 million full-text papers). The searches were performed September 20–29, 2022. SCOPE automatically converted the chemical entities mined in each literature query to SMILES [6] and from these calculated average molecular mass and log *S* and log *P* using the ALogPS3.0 model on the Online CHEmical Modeling (OCHEM) web platform [7]. The results were then visualized as two-dimensional mass and log *P* histograms. SCOPE is written entirely in Python, free, open source and available on GitHub (https://github.com/ReinV/SCOPE) under the Apache 2.0 license.

3. Results

The searches for capillary electrophoresis techniques resulted 1,710,910 occurrences of 13,159 unique chemical entities in 38,798 papers published since 1980, of which only 17 were published in the 1980s, 2828 in the 1990s, 9211 in the 2000s and 19,224 in the 2010s. Unlike the previous static visualizations [2], SCOPE analyses produces interactive visualizations of a chemical 'universe', for example all compounds appearing in the literature on capillary electrophoresis, that can be explored by the user. Closely related compounds are typically found near each other in these maps, and compound classes form distinct shapes in the mass/polarity space. Fig. 1 shows one such view into the chemical universe of capillary electrophoresis. An interactive version of this figure, along with corresponding figures for CAE, CIEF, CITP, CGE, GC, LC, MECC and MEKC for comparison can be found on OSF (https://osf.io/e56zt/).

As in most biomedical literature searches, the two most frequently found chemical entities for both capillary electrophoresis and liquid chromatography are water and glucose. Several of the most frequently found chemicals relate to the equipment, solvent or reagents, rather than the analytes. For example, in the CE search, sodium chloride was found 12,120 times, silicon dioxide/silica 6220 times, and common ions (sodium, lactate, calcium and glutamate) 23,527 times. Conversely, acetonitrile and formic acid were relatively more frequent in the LC literature. None of this is in the least surprising, as CE is typically performed in fused silica capillaries using salt buffers to generate electroosmotic flow. Acetonitrile and formic acid are common in reversed-phase LC, especially when coupled online with mass spectrometry. All of these chemical entities are small, well below 200 Da, and can be suppressed to some extent by term frequency-inverse document frequency normalization in SCOPE, as they are not specific to these analytical methods.

However, when looking at *all* the 9210 chemicals found in the CE literature and for which mass and log *P* could be computed (or 17,236 and 14,390 in the larger LC and GC corpora respectively), it is clear that analytes of larger molecular weight than the salts and solvents dominate the recognized entities (Fig. 1). This is also the region where differences and relative strengths and weaknesses of each method are revealed. Sometimes, SCOPE reveal visually striking and interpretable patterns, originating from over-represented compound classes with certain physicochemical characteristics. For example, the capillary gel electrophoresis analysis show a clear vertical sequence corresponding to nucleotides, which differ in mass, plotted on the vertical axis, but have similar log *P* at −3.5 on the horizontal axis (Fig. 2). Capillary gel (or

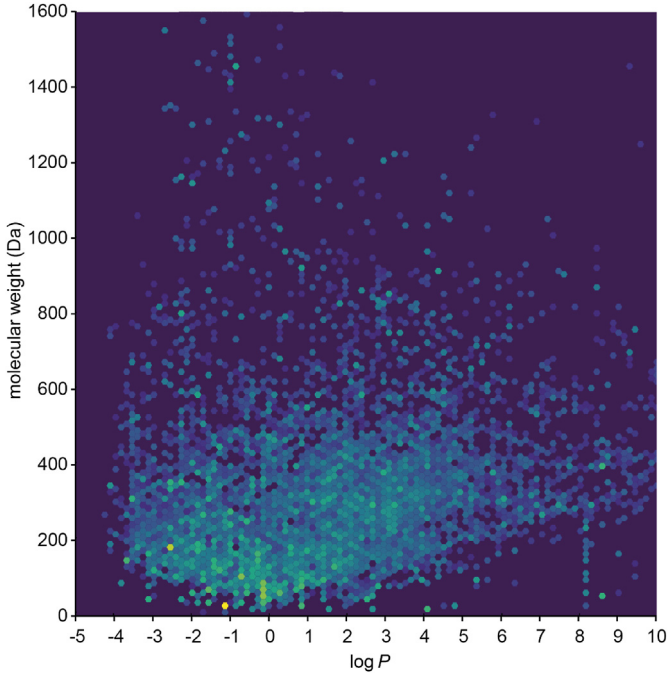

Fig. 1. SCOPE visualization of the capillary electrophoresis literature (saturation 5, blur 0 and TFIDF normalization). Compared to other separation methods, CE has been used to analyze more polar and larger compounds, and rarely for very hydrophobic compounds. The two bright, yellow, bins contain water and glucose respectively.

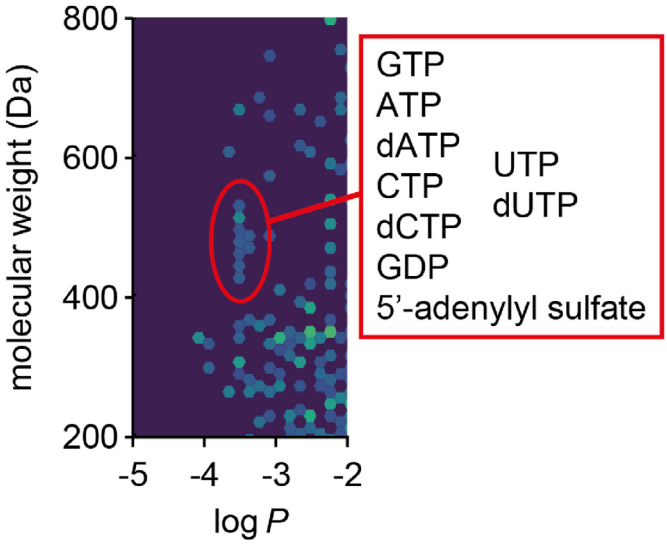

Fig. 2. Region of the capillary gel electrophoresis literature search containing nucleotides, including GTP, dGTP, ATP, dATP, CTP, TTP and 5′-adenylyl sulfate (adenosine-5′-phosphosulfate) with masses from 427 to 523 Da. The two bins at a slightly higher log P (−3.4) contain UTP and dUTP respectively. At similar log P but higher mass we find dinucleotides such as diadenosine 5′,5′-diphosphate and NAD, as well as nucleoside diphosphate sugars such as GDP-α-D-mannose and UDP-α-D-galactose. In the more extensive "capillary electrophoresis" search results, we find many additional nucleotides and nucleosides in this region as well, but they are relatively more frequent in the capillary gel electrophoresis search.

array) electrophoresis was used by many second-generation DNA sequencers [8,9] widely employed in the human genome project in the 1990s [10] and still used in the early 2000s.

Both CIEF and CITP have been used to analyze large, polar molecules, including a number of antibiotics, but CIEF has appears to have seen more application in the analysis of less polar compounds. A compound related to separation modalities that were commonly used in the 1980s, 1990s and early 2000s, but with steadily decreased relative frequency, is the sodium dodecyl sulfate (SDS) used in micellar electrokinetic capillary chromatography (MECC) [11] and for protein separation by capillary gel electrophoresis [12]. Analytes visibly enriched in the MECC and MEKC maps include several glucosides such as baicalin (167 occurrences in 17 papers), catalpol (386 occurrences in 8 papers) and hydroquinone O-β-D-glucopyranoside (149 occurrences in 6 papers) repeatedly analyzed using MECC or MEKC. Other compounds co-occurring with these techniques are caffeine (643 occurrences in 88 papers) and capsaicin (587 occurrences in 10 papers).

Analyzing the statistics reported by SCOPE reveals trends in applicability (Fig. 3). Capillary electrophoresis was applied to larger analytes in the 1990s and 2000s when compared to the 1980s or 2010s (albeit the number of papers from the 1980s is very small), possibly a consequence of a decreasing share of DNA sequencing applications and an increasing popularity of capillary electrophoresis in biomedicine and metabolomics applications. Compared with liquid and gas chromatography, capillary electrophoresis is more often applied to polar analytes, including ionic species, though this difference has decreased in over time (Fig. 3). Though this figure looks simple, it summarizes 22.4 million occurrences of recognized chemical entities (of biological interest) in 450,120 unique publications. In total for all analyses, 36.1 million NERs were retrieved from 589,719 unique publications, 2.0 million of which co-occurring with a capillary electrophoresis technique (including MECC and MEKC) in 42,611 unique publications.

4. Discussion

The results herein demonstrate that it is possible to study the history and application of a particular analytical method or technology by enriching literature searches with a text mining and computational chemistry approach. Advances in ontologies, machine learning and artificial intelligence continually push the state-of-the-art, enabling ever deeper and more specific analyses to answer a wider range of questions. While these analyses are less biased by the investigators' preferences, there are likely still other sources of bias. For example, it is possible that the scientific literature is enriched in reports on novel, challenging or unusual applications of a particular analytical method, and that more routine applications are underreported, obscuring some of the differences in applicability between methods. It is also possible that a technology changes name over time. For example, "capillary zone electrophoresis" being more commonly used in 1980s and 1990s. In more applied work, the underlying technology or method is not always specified. It would therefore be useful to be able to infer these from instrument models (the capillary electrophoresis system, or the DNA sequencer or mass spectrometer). Another caveat is that as the number of publications gets smaller, the effect of a single, prolific, laboratory or research group can no longer be ignored.

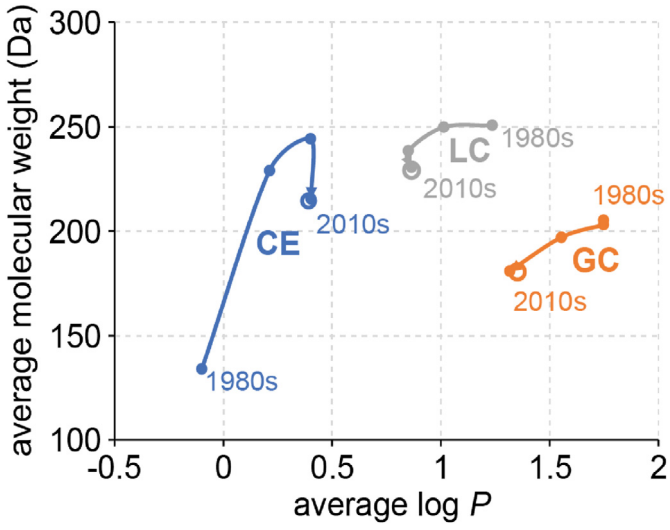

Fig. 3. Trends in applicability of capillary electrophoresis (CE, blue) by decade 1980–2020, compared with liquid chromatography (LC, gray) and gas chromatography (GC, orange) over the same time period. These are the centers of very wide distributions (standard deviations ~200 Da/2.4 log P units for CE and LC and 150 Da/2.9 log P units for GC at all time points). The averages (empty circles) reflect the more recent literature, which dominate in the corpus.

Yet another source of bias is the limitation of the text mining to the small molecules in ChEBI. Peptides, proteins or oligonucleotides are generally not captured, with the exception of peptides with given names. In domains such as proteomics, the analytes, proteins or in bottom-up proteomics peptides resulting from proteolytic digestion, are typically so numerous they are not listed in the papers, but provided as supplemental information or in data repositories. These are not annotated by Europe PMC and were not mined as part of this study. Ideally, perhaps, one would mine the literature for all 112 million compounds in PubChem rather than the 160 thousand in ChEBI, and in all papers in larger databases such as Web of Science. This is technically feasible as long as the calculation of the properties are not too computationally expensive (indeed, we have recalculated the log *P* and log *S* for all compounds in PubChem on the OCHEM server). Nevertheless, SCOPE provides a simple way to get a visual overview of a large body of biochemical literature, and can be generalized to other calculable properties than mass and log *P* or log *S*, such as collisional cross sections relevant in the context of ion mobility separations, sometimes referred to as gaseous electrophoresis [13]. Ion mobility was combined with capillary electrophoresis already in 1989 by Hallen et al. [14], and has recently been used to analyze low-nanogram samples in proteomics [15]. Biological properties such as toxicity or blood-brain barrier permeability may be also computed and visualized in SCOPE.

5. Conclusions

To our knowledge, this is the first time that text mining and computational chemistry have been combined to analyze applications and applicability of capillary electrophoresis. While the most frequently occurring compounds in the literature relate to capillaries and buffers, the bulk of the annotated chemicals relate to the analytes. These differ between analytical techniques, revealing strengths and weaknesses of each. Importantly, literature searches in SCOPE could easily be refined by restricting the search to a certain time period, journal, country or domain of application, focusing the investigation into a particular aspect of the topic.

Author contributions

Magnus Palmblad: Conceptualization, Methodology, Formal analysis, Data curation, Writing- Original draft preparation, Supervision. Reiner Vleugels: Data curation, Methodology, Software, Reviewing and Editing. Jonas Bergquist: Validation, Writing-Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data is publicly available on OSF, including the results presented in this manuscript.

Acknowledgements

The Europe PMC team is gratefully acknowledged for their support using programmatic access to the literature databases and Dr. Igor Tetko is equally acknowledged for assistance using the OCHEM servers.

References

[1] R.L.C. Voeten, I.K. Ventouri, R. Haselberg, G.W. Somsen, Capillary electrophoresis: trends and recent advances, Anal. Chem. 90 (2018) 1464.
[2] M. Palmblad, Visual and semantic enrichment of analytical chemistry literature searches by combining text mining and computational chemistry, Anal. Chem. 91 (2019) 4312.
[3] R. Vleugels, M. Palmblad, Non-Uniform Gaussian Blur of Hexagonal Bins in Cartesian Coordinates (Preprint), 2020. https://doi.org/10.48550/arXiv.2005.09941. arXiv.
[4] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, C. Steinbeck, The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013, Nucleic Acids Res. 41 (2013) D456.
[5] A. Venkatesan, J.H. Kim, F. Talo, M. Ide-Smith, J. Gobeill, J. Carter, R. Batista-Navarro, S. Ananiadou, P. Ruch, J. McEntyre, SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data, Wellcome Open Res 1 (2016) 25.
[6] E. Anderson, G.D. Veith, D. Weininger, SMILES: a Line Notation and Computerized Interpreter for Chemical Structures, 1987.
[7] I. Sushko, S. Novotarskyi, R. Korner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V.V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, Baskin II, V.A. Palyulin, E.V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I.V. Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, J. Comput. Aided Mol. Des. 25 (2011) 533.
[8] H. Swerdlow, R. Gesteland, Capillary gel electrophoresis for rapid, high resolution DNA sequencing, Nucleic Acids Res. 18 (1990) 1415.
[9] J.M. Heather, B. Chain, The sequence of sequencers: the history of sequencing DNA, Genomics 107 (2016) 1.
[10] B.L. Karger, A. Guttman, DNA sequencing by CE, Electrophoresis 30 (Suppl 1) (2009) S196.
[11] S. Terabe, K. Otsuka, K. Ichikawa, A. Tsuchiya, T. Ando, Electrokinetic separations with micellar solutions and open-tubular capillaries, Anal. Chem. 56 (1984) 111.
[12] A.S. Cohen, B.L. Karger, High-performance sodium dodecyl sulfate polyacrylamide gel capillary electrophoresis of peptides and proteins, J. Chromatogr. 397 (1987) 409.
[13] H.E. Revercomb, E.A. Mason, Theory of plasma chromatography gaseous electrophoresis - review, Anal. Chem. 47 (1975) 970.
[14] R.W. Hallen, C.B. Shumate, W.F. Siems, T. Tsuda, H.H. Hill Jr., Preliminary investigation of ion mobility spectrometry after capillary electrophoretic introduction, J. Chromatogr. 480 (1989) 233.
[15] K.R. Johnson, M. Gregus, A.R. Ivanov, Coupling high-field asymmetric ion mobility spectrometry with capillary electrophoresis-electrospray ionization-tandem mass spectrometry improves protein identifications in bottom-up proteomic analysis of low nanogram samples, J. Proteome Res. 21 (2022) 2453.