Hinselmann *et al. Journal of Cheminformatics* 2010, **2**(Suppl 1):P36
http://www.jcheminf.com/content/2/S1/P36

Journal of Cheminformatics

POSTER PRESENTATION · Open Access

Efficient extraction of canonical spatial relationships using a recursive enumeration of k-subsets

Georg Hinselmann*, Nikolas Fechner, A Jahn, Andreas Zell

From 5th German Conference on Cheminformatics: 23. CIC-Workshop
Goslar, Germany. 8-10 November 2009

The spatial arrangement of a chemical compound plays an important role regarding the related properties or activities. A straightforward approach to encode the geometry is to enumerate pairwise spatial relationships between k substructures, like functional groups or subgraphs. This leads to a combinatorial explosion with the number of features of interest and redundant information. The goal of this work is to compute all possible k-subsets of spatial points and to extract a single canonical descriptor for each subset in sub-polynomial computation time. More precisely, the problem is to reduce the complexity of $n_k = n \cdot (n - 1)...(n - k)$ possible relationships (patterns or descriptors) for n features and k-point relationships.

We propose a two-step algorithm to solve this problem. A modified algorithm for the computation of the binomial coefficient computes the k-subsets [1] containing the possible combinations of the n relevant features. If a k-subset is completed in the inner recursion, the algorithm computes a canonical representation for it. By defining a natural order by means of the geometrical center of gravity of the k points, we extract k patterns that describe the distance to the center of gravity and type of the spatial feature $k \in F$. Then, the algorithm returns a unique identifier for the lexicographically sorted array of patterns. If applicable ($f_{\pi_0} < f_{\pi_1} < \dots f_{\pi_k}$), an additional identifier is added which has the form $f_{\pi_o} \xrightarrow{d_{\pi_0\pi_1}} f_{\pi_1} \xrightarrow{\dots} \dots f_{\pi_{k-1}} \xrightarrow{d_{\pi_{k-1}\pi_k}} f_{\pi_k}$, where d_{ij} denotes the geometrical distance between features i, j. Else ($f_{\pi_o} \le f_{\pi_1} \le \dots f_{\pi_k}$), this step is omitted. Therefore, this approach also considers stereochemistry. Finally, one feature is returned for each k-subset resulting in a set of $C(n, k)$ patterns describing the structure.

The main result is that the number of features is reduced from n_k to $C(n, k)$, which equals the binomial coefficient. This procedure is useful in combination with similarity approaches that use spatial relationships, like pharmacophore searches, fingerprints, or graph kernels. We experimentally validated the algorithm on numerous QSAR benchmark sets in combination with the pharmacophore kernel [2].

Published: 4 May 2010

References

1. Rolfe T: *SIGCSE Bull* 2001, **33(3)**:35-36.
2. Mahé P, Ralaivola L, Stoven V, Vert J-P: *J Chem Inf Mod* 2006, **46(5)**:2003-2014.

doi:10.1186/1758-2946-2-S1-P36
Cite this article as: Hinselmann *et al.*: Efficient extraction of canonical spatial relationships using a recursive enumeration of k-subsets. *Journal of Cheminformatics* 2010 **2**(Suppl 1):P36.

University of Tübingen, Sand 1, 72076 Tübingen, Germany