JOURNAL OF SYSTEMS CHEMISTRY

RESEARCH ARTICLE Open Access

Systems chemistry: using thermodynamically controlled networks to assess molecular similarity

Vittorio Saggiomo[1†], Yana R Hristova[2†], R Frederick Ludlow[2,3] and Sijbren Otto[1*]

Abstract

Background: The assessment of molecular similarity is a key step in the drug discovery process that has thus far relied almost exclusively on computational approaches. We now report an experimental method for similarity assessment based on dynamic combinatorial chemistry.

Results: In order to assess molecular similarity directly in solution, a dynamic molecular network was used in a two-step process. First, a clustering analysis was employed to determine the network's innate discriminatory ability. A classification algorithm was then trained to enable the classification of unknowns. The dynamic molecular network used in this work was able to identify thin amines and ammonium ions in a set of 25 different, closely related molecules. After training, it was also able to classify unknown molecules based on the presence or absence of an ethylamine group.

Conclusions: This is the first step in the development of molecular networks capable of predicting bioactivity based on an assessment of molecular similarity.

Keywords: Dynamic combinatorial chemistry, Systems chemistry, Molecular networks, Data mining, Clustering analysis

Background

Molecular similarity relates to the extent to which molecules have similar structures or properties. Hence, molecular similarity and any quantification of it are both strongly context dependent. Assessing molecular similarity is a key element in the drug discovery process as structural similarity is believed to be correlated to activity with respect to a given target [1-4]. However, assessing molecular similarity is not trivial. The most common approaches involve computational methods, including the use of molecular fingerprints [2], simple calculated properties such as solvent accessible surface area, number of hydrogen-bond donor and acceptor groups, etc. [5-7] or shape comparisons [8-10]. Three-dimensional methods, such as CoMFA [11] and CoMSIA [12], map favorable and unfavorable interaction regions around or onto the structure of a molecule, requiring prior knowledge of the appropriate conformations of this molecule.

We reasoned that a realistic measure of molecular similarity may be obtained by interrogating the molecules in solution experimentally. The closest to an experimental approach to analysing molecular similarity are sensing systems, where the objective is usually the detection and quantification of a specific analyte or the discrimination between different analytes. Such assays have been set up in array format [13-16] and more recently also using dynamic combinatorial chemistry [17,18]. However, we are not aware of any examples of the use of these approaches for determining similarity.

We now report the adaptation of dynamic combinatorial chemistry for similarity assessment. The central premise of our approach is that the extent of binding of a molecule by a synthetic receptor contains information about the structure of the molecule. While binding by a single receptor will provide only very limited information, a more comprehensive description of the molecular structure may be obtainable by using a systems chemistry [19-24] approach, utilising the binding to multiple receptors. Specifically, we employed a dynamic molecular network containing a variety of potential synthetic receptors. These receptors are connected through reversible

* Correspondence: s.otto@rug.nl
†Equal contributors
[1]Centre for Systems Chemistry, Stratingh Institute, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands
Full list of author information is available at the end of the article

covalent bonds and therefore continuously exchange their constituent building blocks. Through work on dynamic combinatorial libraries [25-29], it is well established that dynamic molecular networks will change their composition in response to molecular recognition by an introduced effector, leading to a redistribution of the building blocks in favor of those receptors with affinity for the effector. This effect has so far mainly been exploited as a tool for identifying individual receptors and for constructing sensor networks [30-35]. We now show how such a network can make a rudimentary assessment of molecular similarity. In this approach there is no need to synthesise all the receptors separately; they are generated in one step when preparing the dynamic combinatorial library. Yet it is possible to identify the individual receptors in the mixture using LC-MS.

Results and discussion

We selected a set of amines and ammonium ions (**4**-**28**), shown in Figure 1, as effector molecules, featuring functional groups that are common in many drugs. We constructed the dynamic molecular network of potential receptors from dithiol building blocks **1**-**3**. These building blocks feature carboxylic acid groups that can potentially recognise the amine and ammonium groups of the effectors through hydrogen bonding and electrostatic interactions. They also contain aromatic rings that may engage in hydrophobic interactions with the set of effector molecules. Each building block features two thiol groups, which can be oxidised to disulfides, giving rise to a mixture of macrocycles that can equilibrate through disulfide exchange [36-38].

Thus, exposing an equimolar solution of **1**-**3** (5 mM total) in borate buffer (50 mM, pH 8.0) to atmospheric oxygen for three days gave a mixture of disulfide macrocycles dominated by $(\mathbf{1})(\mathbf{3})_2$, $(\mathbf{3})_4$, $(\mathbf{3})_3$, $(\mathbf{1})(\mathbf{2})(\mathbf{3})$, $(\mathbf{1})_2(\mathbf{3})$ and $(\mathbf{1})_2(\mathbf{2})_2$. We analysed the response of this small molecular network to the introduction of the individual effectors (2.5 mM) by LC-MS (Representative chromatograms are shown in Figure 2).

We determined the amplification factors (i.e. the ratio of the HPLC peak areas in the presence and absence of

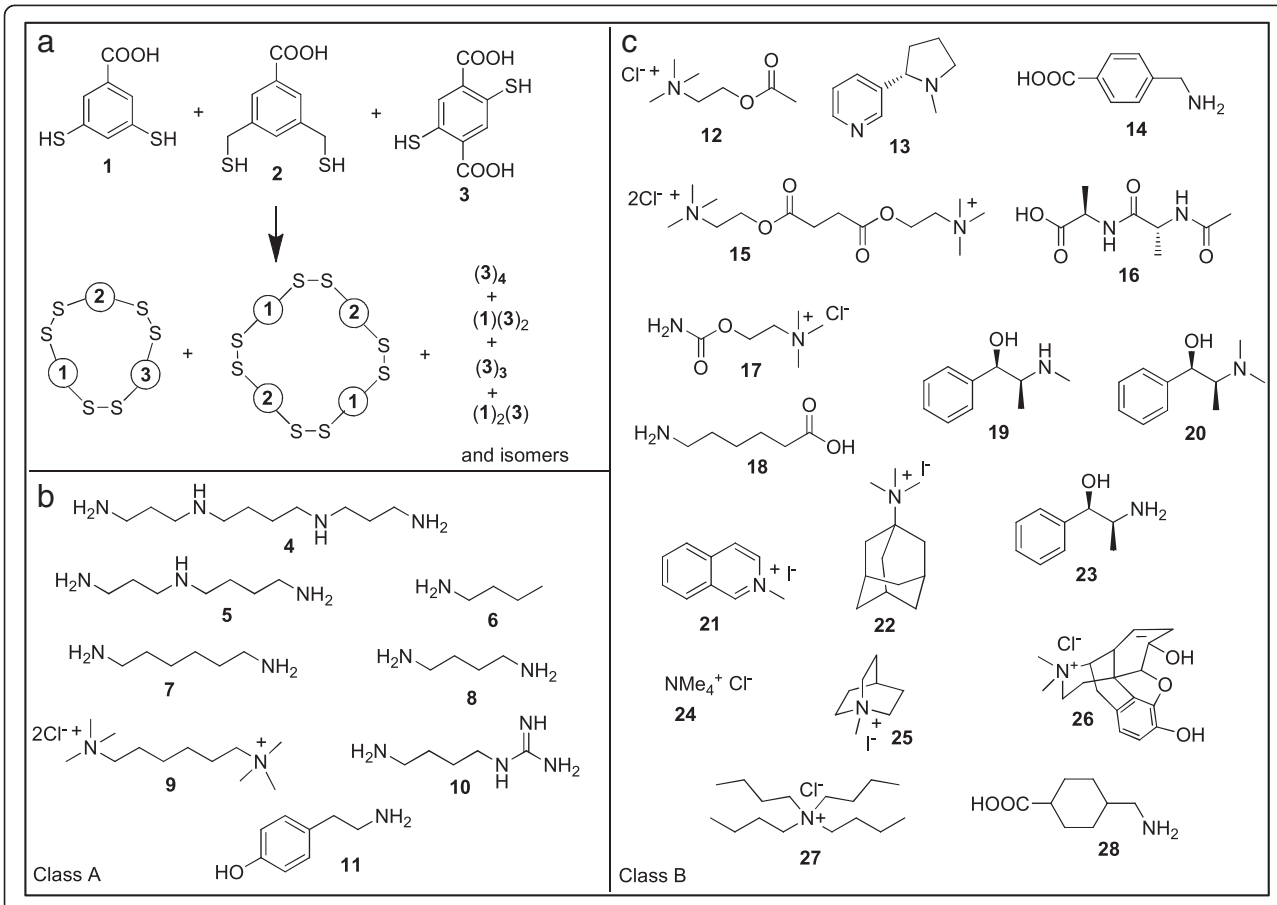

Figure 1 a) Dithiol building blocks 1-3 are oxidised to give a dynamic network consisting predominantly of trimeric and tetrameric potential receptors. A clustering analysis of the response of the receptor network to a set of effectors separates them into two groups: class **A** (**b**) and class **B** (**c**).

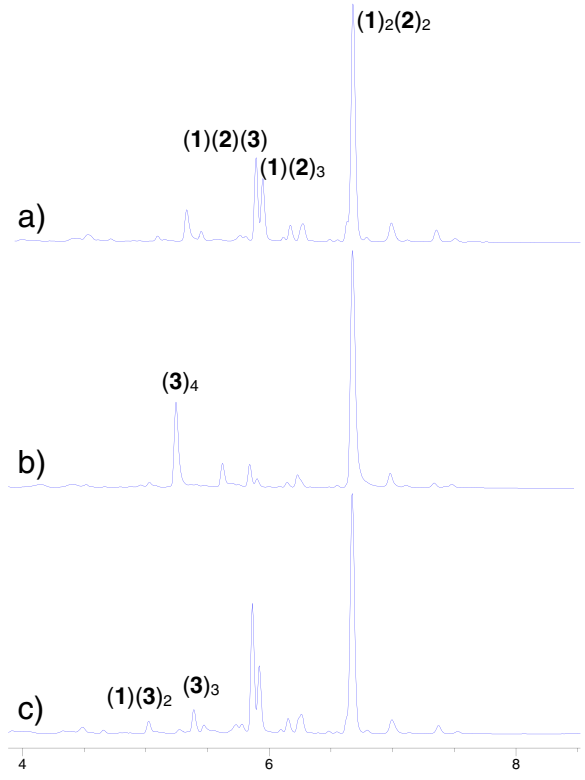

Figure 2 HPLC analysis of a) the dynamic network made from an equimolar solution of building blocks 1-3 (5 mM total) in borate buffer (50 mM, pH 8.0) without any effector; b) in presence of effector 5 (class A) and c) in presence of effector 15 (class B).

the effectors) for the six dominant network members and for all 25 effector molecules (Additional file 1).

The approach we took for investigating the ability of the network to assess molecular similarity is summarised in Figure 3. We first probed the network's innate ability to discriminate between different classes of effectors using a clustering analysis. We then used the thus uncovered classification into clusters to train a classification algorithm. Finally, we tested the performance of the dynamic molecular network by challenging it with "unknowns".

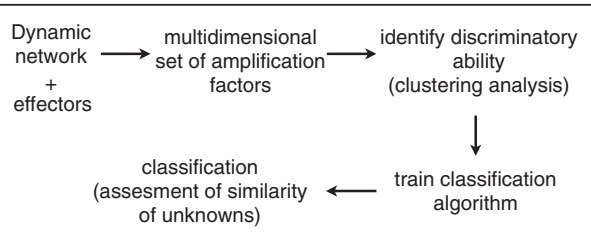

Figure 3 A dynamic network responds differently to different effectors producing a multidimensional dataset of amplification factors. These are used in a clustering analysis that shows the innate discriminatory ability of the network. The same data is then used to train an algorithm that will allow the classification of unknowns.

In a clustering analysis a given set of measurements is divided into two or more clusters based only on the distances of all the points in *n*-dimension, where *n* is the number of variables. Without requiring any transformation of the original dataset this analysis iteratively finds the centroids of two or more clusters, following two main rules: a centroid must be close to the largest possible number of points and, at the same time, must be far away from the other centroid(s). We used an unsupervised method, i.e. classes were assigned autonomously during the analysis without requiring any user input. The dataset composed of the amplification factors of the various receptors upon addition of different effector molecules **4**-**28** was subjected to a k-means clustering analysis [39]. K-means is a partitioning algorithm with a chosen number of cluster centroids k, that tries to minimise the sum of within-cluster-variances. Each object is assigned to a k midpoint on the basis of Euclidean distance [40]. This k midpoint is then recalculated based on the average of all points assigned to it. These processes are iteratively repeated until each k is at the centre of the cluster. The number of centroids starts at two and increases until a cluster with only one or two points is found. In our case already the third cluster had only two points. Two main clusters were identified: class A (consisting of effectors **4**-**11**) and class B (consisting of effectors **12**-**28**). Figure 4 shows a graphical representation of those clusters reduced to only two dimensions.

Inspection of the nature of the clustered molecules revealed that the network has an innate ability to discriminate the relatively thin amines and ammonium ions from a range of different amines and ammonium ions that are either more bulky or carry negative (partial) charge (Figure 1). Having established the discriminatory ability of the network, we investigated whether we could use the network for the classification of "unknown" molecules. More specifically, we investigated the possibility of using the network's response to predict whether molecules contain the ethylamine group. Our network seemed highly suitable for this, since, with the exception of effector **9**, all molecules in class A contain an ethylamine group, while, with the exception of **18**, none of those in class B do.

In a classification analysis (supervised learning) unknown objects are classified based on the comparison of their variables with those of a training set with predefined classes. We opted for the use of the naïve Bayes classifier [41]. The naïve Bayes is a simple probabilistic classifier that requires a small amount of training data to estimate the parameters for the comparison. All the variables contribute independently to the assignment of an unknown object to a predetermined class. For this analysis a

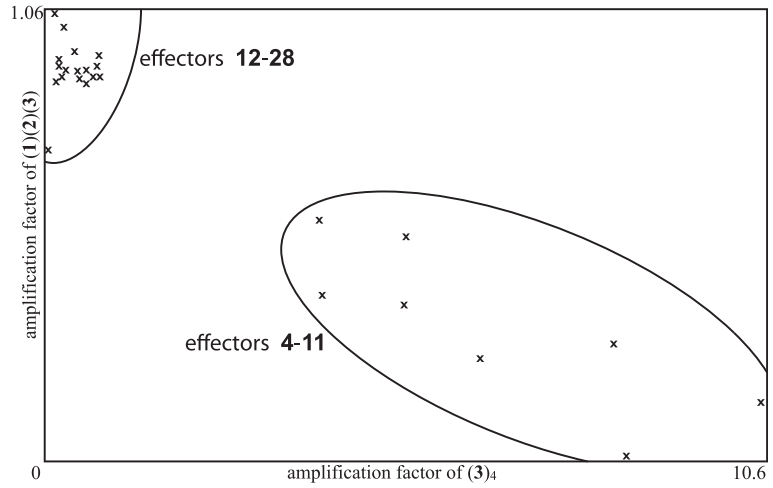

Figure 4 The k-means clustering analysis resulted in the assignment of the effectors 4-28 to two different clusters, based on the response of the addition of these effectors to a molecular network made from building blocks 1-3. This graph is a representation of the two clusters in only two dimensions, while the clustering analysis was performed using all six dimensions.

training dataset was set up with all the amplification factors of all effectors except one, used as unknown (**4-8**, **10-11** and **18** for class A and effectors **9**, **12-17** and **19-28** for class B). Then the amplification factors of the unknown effector were subjected to naïve Bayesian classification analysis in Weka [42] for the class assignment. In 23 out of 25 cases the unknown was assigned to the right class (92% correct assignment). Only effectors **9** and **18** were wrongly assigned, the first was assigned to class A while the second to class B, as in the clustering experiment. This cross validation experiment establishes that the molecular network, when properly trained, is able to classify unknowns.

Conclusions

We have shown how a simple molecular network can perform a rudimentary assessment of molecular similarity and can successfully classify unknowns. To the best of our knowledge this is the first experimental approach to assess molecular similarity and these results represent the first step towards developing networks that may be able to discriminate and assess similarity of biologically active molecules and drugs and potentially predict bioactivity. However, there is still a long road ahead. In the present clustering approach the similarity that is assessed is dictated by the innate discriminatory ability of the network, and is only known after the response of the molecular network to a series of effectors has been analysed. Many more such studies on different dynamic networks are needed before we will be able to design molecular networks that will perform well in clustering molecules based on a pre-defined similarity parameter. In contrast, a classification analysis may yield useful results more readily, as the scientist can decide the parameter on which the classification should be based, whereafter the algorithm selects the data that is most discriminatory for this particular parameter. We are currently working towards this vision by using more molecular networks that exhibit increased structural diversity.

Methods

General methods

Building blocks **1** [37], **2** [43], **3** [44], and effector **22** [45] were synthesised following literature procedures. All other effectors were obtained from commercial sources and used without further purification. HPLC analysis was performed on Agilent 1050 or 1100 systems coupled to a UV detector. LC-MS analysis was performed using an Agilent XCT ion trap MSD mass spectrometer. Mass spectra (negative ion mode) were acquired in ultra-scan mode using a drying temperature of 350°C, a nebuliser pressure of 35.00 psi, drying gas flow of 9 L/min, capillary voltage 4000 V and an ICC target of 10,000 ions. Agilent Chemstation software (Rev A.10.02) and Bruker Daltonik LC/MSD Trap software 5.2 (Build 374) was used to operate the LC-MS and analyse the data. For the LC and LC-MS a Zorbax XDB-C8, 2.1 × 150 mm column was used at 40°C with a gradient (flow rate 0.2 mL/min) from 5% to 95% of acetonitrile in water (both solvents containing 0.1% of formic acid).

Dynamic network preparation and analysis

Building blocks **1**-**3** were dissolved together in a 50 mM borate buffer solution (pH 8.0) with a total final concentration of 5 mM. Effectors were added separately at a concentration of 2.5 mM. The libraries were stirred for 3 days and then analysed by HPLC and LC-MS.

Statistical methods

Weka (GNU GPL) ver. 3.7.1 was used on Mac OSX. A text file with all amplification factors of each effector was used as input file for the k-means clustering analysis with Weka using Euclidian distances applying the parameters: "weka. clusterers.SimpleKMeans-N2-A "weka.core.EuclideanDistance - R first-last" -I 500 -S 10". The same file except one effector was used as training set for the naïve Bayes classification analysis. The effector removed from the training set input data was used as unknown. The standard parameters used for this analysis were: "weka. classifiers.bayes.NaiveBayes".

Additional file

Additional file 1: Supporting information contains the effector-induced amplification factors of the six receptors in the molecular network.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

R.F.L. conceived the idea, Y.R.H. prepared and analysed the molecular networks. V.S. performed the data analysis, analysed the results and wrote the paper with S.O. All authors read and approved the final manuscript.

Acknowledgements

We thank EPSRC, COST CM0703, COST CM1005, Marie Curie RTN DCC, The Netherlands Organization for Scientific Research (NWO) (V.S.) and P.T. Corbett for useful discussions.

Author details

[1]Centre for Systems Chemistry, Stratingh Institute, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands. [2]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom. [3]Present address: Astex Pharmaceuticals, 436 Cambridge Science Park, Cambridge CB4 0QA, United Kingdom.

Received: 7 December 2012 Accepted: 23 January 2013
Published: 12 February 2013

Reference

1. Horvath D, Jeandenans C: **Neighborhood Behavior of In Silico Structural Spaces with respect to In Vitro Activity Spaces – A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles.** *J Chem Inf Comput Sci* 2003, **43:**680–690.
2. Bender A, Glen RC: **Molecular Similarity: a Key Technique in Molecular Informatics.** *Org Biomol Chem* 2004, **2:**3204–3218.
3. Boström J, Hogner A, Schmitt S: **Do Structurally Similar Ligands Bind in a Similar Fashion?** *J Med Chem* 2006, **49:**6716–6725.
4. Eckert A, Bajorath J: **Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations, and novel Approaches.** *Drug Discov Today* 2007, **12:**225–233.
5. Martin EJ, Blaney JM, Siani MA, Spellmeyer DC, Wong AK, Moos WH: **Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery.** *J Med Chem* 1995, **38:**1431–1436.
6. Amin EA, Welsh WJ: **A preliminary in Silico Lead Series of 2-phthalimidinoglutaric Acid Analogues Sesigned as MMP-3 Inhibitors.** *J Chem Inf Model* 2006, **46:**2104–2109.
7. Jennings A, Tennant M: **Selection of Molecules Based on Shape and Electrostatic Similarity: Proof of Concept of "Electroforms".** *J Chem Inf Model* 2007, **47:**1829–1838.
8. Grant JA, Gallardo MA, Pickup B: **A Fast Method of Molecular Shape Comparison: a Simple Application of a Gaussian Description of Molecular Shape.** *J Comput Chem* 1996, **17:**1653–1666.
9. Rush TS, Grant JS, Mosyak L, Nicholls A: **A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein–Protein Interaction.** *J Med Chem* 2005, **48:**1489–1495.
10. Ballester PG, Richards WG: **Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes.** *J Comput Chem* 2007, **28:**1711–1723.
11. Cramer RD III, Patterson DE, Bunce JD: **Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroids to Carrier Proteins.** *J Am Chem Soc* 1988, **110:**5959–5967.
12. Klebe G, Abraham U, Mietzner T: **Molecular Similarity Indexes in a Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict their Biological Activity.** *J Med Chem* 1994, **37:**4130–4146.
13. Janowski V, Severin K: **Carbohydrate Sensing with a Metal-Based Indicator Displacement Assay.** *Chem Commun* 2011, **47:**8521–8523.
14. Shabbir SH, Joyce LA, DeCruz GM, Lynch VM, Sorey S, Anslyn EV: **Pattern-Based Recognition for the Rapid Determination of Identity, Concentration and Enantiomeric Excess of Subtly Different Diols.** *J Am Chem Soc* 2009, **131:**13125–13131.
15. Hewage HS, Anslyn EV: **Pattern-Based Recognition of Thiols and Metals Using a Single Squarane Indicator.** *J Am Chem Soc* 2009, **131:**13099–13106.
16. Nguyen BT, Anslyn EV: **Indicator- Displacement Assays.** *Coord Chem Rev* 2005, **250:**3118–3127 and refs therein.
17. Rochat S, Severin K: **Pattern-Based Sensing with Metal–Dye Complexes: Sensor Arrays versus Dynamic Combinatorial Libraries.** *J Comb Chem* 2010, **12:**595–599.
18. Montenegro J, Bonvin P, Takeuchi T, Matile S: **Dynamic Octopus Amphiphiles as Powerful Activators of DNA Transporters: Differential Fragrance Sensing and Beyond.** *Chem Eur J* 2010, **16:**14159–14166.
19. Whitesides GM, Ismagilov RF: **Complexity in Chemistry.** *Science* 1999, **284:**89–92.
20. Ludlow RF, Otto S: **Systems Chemistry.** *Chem Soc Rev* 2008, **37:**101–108.
21. Peyralans JJP, Otto S: **Recent Highlights in Systems Chemistry.** *Curr Opin Chem Biol* 2009, **13:**705–713.
22. Nitschke JR: **Systems Chemistry: Molecular Networks Come of Age.** *Nature* 2009, **462:**736–738.
23. Gibb BC: **Teetering Towards Chaos and Complexity.** *Nat Chem* 2009, **1:**17–18.
24. von Kiedrowski G, Otto S, Herdewijn P: **Welcome Home, Systems Chemists!** *J Syst Chem* 2010, **1:**1–16.
25. Corbett PT, Leclaire J, Vial L, West KR, Wietor J-L, Sanders JKM, Otto S: **Dynamic Combinatorial Chemistry.** *Chem Rev* 2006, **106:**3652–3711.
26. Ladame S: **Dynamic Combinatorial Chemistry: on the Road to Fulfilling the Promise.** *Org Biomol Chem* 2008, **6:**219–226.
27. Reek JHR, Otto S: *Dynamic Combinatorial Chemistry*. Weinheim: Wiley-VCH; 2010.
28. Miller BL: *Dynamic Combinatorial Chemistry An Introduction, in Dynamic Combinatorial Chemistry: In Drug Discovery, Bioorganic Chemistry, and Materials Science.* Hoboken: Wiley & Sons; 2010.
29. Hunt RAR, Otto S: **Dynamic Combinatorial Libraries: New Opportunities in Systems Chemistry.** *Chem Commun* 2011, **47:**847–858.
30. Besenius P, Cormack PAG, Ludlow RF, Otto S, Sherrington DC: **Affinity Chromatography in Dynamic Combinatorial Libraries: One-Pot Amplification and Isolation of a Strongly Binding Receptor.** *Org Biomol Chem* 2010, **8:**2414–2418.
31. Klein JM, Saggiomo V, Reck L, McPartlin M, Dan Pantoş G, Lüning U, Sanders JKM: **A Remarkably Flexible and Selective Receptor for Ba2+ Amplified from a Hydrazone Dynamic Combinatorial Library.** *Chem Commun* 2011, **47:**3371–3373.
32. Buryak A, Pozdnoukhov A, Severin K: **Pattern-Based Sensing of Nucleotides in Aqueous Solution with a Multicomponent Indicator Displacement Assay.** *Chem Commun* 2007, **23:**2366–2368.
33. Buryak A, Zaubitzer F, Pozdnoukhov A, Severin K: **Indicator Displacement Assays as Molecular Timers.** *J Am Chem Soc* 2008, **130:**11260–11261.
34. Zaubitzer F, Riis-Johannessen T, Severin K: **Sensing of Peptide Hormones with Dynamic Combinatorial Libraries of Metal–Dye Complexes: the Advantage of Time-Resolved Measurements.** *Org Biomol Chem* 2009, **7:**4598–4603.
35. Montenegro J, Fin A, Matile S: **Comprehensive Screening of Octopus Amphiphiles as DNA Activators in Lipid Bilayers: Implications on Transport, Sensing and Cellular Uptake.** *Org Biomol Chem* 2011, **9:**2641–2647.

36. Otto S, Furlan RLE, Sanders JKM: **Dynamic Combinatorial Libraries of Macrocyclic Disulfides in Water.** *J Am Chem Soc* 2000, **122:**12063–12064.
37. Otto S, Furlan RLE, Sanders JKM: **Selection and Amplification of Hosts from Dynamic Combinatorial Libraries of Macrocyclic Disulfides.** *Science* 2002, **297:**590–593.
38. West K, Bake K, Otto S: **Dynamic Combinatorial Libraries of Disulfide Cages in Water.** *Org Lett* 2005, **7:**2615–2618.
39. Witten IH, Frank E: *"Iterative distance-based clustering" in Data Mining*. 2nd edition. San Francisco: Elsevier; 2005:137–138.
40. The Euclidean distance of two points is defined as the length of the line segment connecting them.
41. Witten IH, Frank E: *"Clustering for classification" in Data Mining*. 2nd edition. San Francisco: Elsevier; 2005:337–338.
42. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA Data Mining Sofware: an Update.** *SIGKDD Explorations* 2009, **11:**10–19.
43. Staab HA, Kirrstetter RGH: **[2.2](2,7)Pyrenophan als Excimeren-Modell: Synthese und Spektroskopische Eigenschaften.** *Liebigs Ann Chem* 1979, 886–898.
44. Vial L, Ludlow RF, Leclaire J, Pérez-Fernández R, Otto S: **Controlling the Biological Effects of Spermine Using a Synthetic Receptor.** *J Am Chem Soc* 2006, **128:**10253–10257.
45. Kondo Y, Uematsu R, Nakamura Y, Kusabayashi S: **Empirical Analysis on the Constituent Terms of Transfer Enthalpies.** *J Chem Soc Faraday Trans 1* 1988, **84:**111–116.

doi:10.1186/1759-2208-4-2
Cite this article as: Saggiomo *et al.*: Systems chemistry: using thermodynamically controlled networks to assess molecular similarity. *Journal of Systems Chemistry* 2013 **4**:2.