PEDS Advance Access published June 8, 2007

Protein Engineering, Design & Selection pp. 1–5, 2007
doi:10.1093/protein/gzm027

Amino acid quantitative structure property relationship database: a web-based platform for quantitative investigations of amino acids

Yi Lu[1], Blazej Bulka[2], Marie desJardins[2] and Stephen J. Freeland[1,3]

Departments of [1]Biological Sciences and [2]Computer Science and Electrical Engineering, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

[3]To whom correspondence should be addressed. E-mail: freeland@umbc.edu

Here, we present the AA-QSPR Db (Amino Acid Quantitative Structure Property Relationship Database): a novel, freely available web-resource of data pertaining to amino acids, both engineered and naturally occurring. In addition to presenting fundamental molecular descriptors of size, charge and hydrophobicity, it also includes online visualization tools for users to perform instant, interactive analyses of amino acid sub-sets in which they are interested. The database has been designed with extensible markup language technology to provide a flexible structure, suitable for future development. In addition to providing easy access for queries by external computers, it also offers a user-friendly web-based interface that facilitates human interactions (submission, storage and retrieval of amino acid data) and an associated e-forum that encourages users to question and discuss current and future database contents.

Keywords: amino acids/database/QSPR/XML

Introduction

Beyond protein synthesis, amino acids play many significant roles in biology: as intermediates of metabolic pathways, neurotransmitters, antibiotics, etc. Furthermore, amino acids that have been never synthesized in nature are now routinely incorporated into biological systems to aid scientists who investigate fundamental questions of biology and medicinal chemistry. Thus, quantitative investigations of the relationship between amino acids, natural and engineered, are not only critical for biological research and bioengineering (such as predicting the biological activity of natural and engineered peptides, e.g. Guan *et al.*, 2005) but are also informative for investigating the origin and evolution of early life (Lu and Freeland, 2006a). However, although extensive biochemical, bioinformatic and evolutionary studies of the 20 'standard' (biologically encoded) amino acids that polymerize to make proteins have led to the creation of associated web resources, (e.g. Kawashima *et al.*, 1999 and the NCBI's amino acid explorer: http://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi), corresponding data and analysis tools are scarce for the hundreds of amino acids that form a chemical context around them.

This disparity of information between the 20 biologically encoded amino acids and all others has been in large part attributable to the costly and time-consuming traditional experimental approach required to measure amino acid biophysical properties (Haidacher *et al.*, 1996). However, recent developments in computational chemistry offer us an alternative method to quickly and reliably estimate values for key amino acid properties (for example, freely accessible web software can predict van der Waals volume with an accuracy, measured as coefficient of determination between predicted and experimentally determined values, of 0.955; Lu and Freeland, 2006b).

Therefore, here we introduce a novel and freely accessible online extensible markup language (XML) database, the AA-QSPR Db. Currently, the database comprises a total of 388 amino acids: the 20 amino acids found in the standard genetic code, 177 that have been found in biological systems, acting as intermediates in main metabolic pathways, neurotransmitters (Venton *et al.*, 2006) and antibiotics (Czajgucki *et al.*, 2006) but have never been incorporated into the genetic code, for example, ornithine and sarcosine (Garrett and Grisham, 1999), 69 that are thought to be synthesized abiotically (Cronin and Pizzarello, 1986), 108 that are products of post-translational modification (Uy and Wold, 1977) and 58 that have been engineered by scientists (Summerer *et al.*, 2006).

The abiotic amino acids are mainly identified from analyses of the Murchison meteorite (Cronin *et al.*, 1981, 1985; Cronin and Pizzarello, 1986), or described as products of pre-biotic simulation experiments (Miller, 1986) and are particularly relevant to thinking in exobiology and the origin of life (Cronin and Pizzarello, 1997; Glavin and Bada, 2001). Engineered amino acids are largely drawn from recent progress in the research of the 'incorporation of nonnatural amino acids into proteins' (Link *et al.*, 2003; Hendrickson *et al.*, 2004; Wang *et al.*, 2006).

Our AA-QSPR Db comes with an associated toolkit that provides broad utility within the field of biochemical ontology: possible applications range from research into protein structure and amino acid bioactivity to synthetic biology and evolutionary analyses.

Technical description of the database

Contents and tools of the database

Amino acids considered A major criterion we used to determine biochemical relevance of an amino acid (and thus its incorporation into our database) is its ability to form peptide bond with another amino acid. Thus, we did not include 2,5-diaminopyrrole, an amino acid derivative found in Murchison meteorite (Meierhenrich *et al.*, 2004), because it has no free carboxyl group and is thus of limited interest to research concerning biological macromolecules. However, we did not limit database contents to amino acids that possess an alpha-amino group and a free alpha-hydrogen.

© The Author 2007. Published by Oxford University Press. All rights reserved.
For Permissions, please e-mail: journals.permissions@oxfordjournals.org

Although these are two features shared by all the 20 standard amino acids (Weber and Miller, 1981), they may not be logical constraints on biochemistry (Qiu *et al.*, 2006), either in terms of primordial evolution or in terms of protein engineering.

Biophysical properties considered Within our database, each amino acid is currently associated with quantitative estimates of three fundamental biophysical properties: size, charge and hydrophobicity. These three amino acid properties have long been regarded as major determinants of amino acids' bioactivity (Grantham, 1974; Biro, 2006), influencing not only the biochemical roles played by amino acids, but also the patterns of molecular evolution that occur and hence the expectations of bioinformatics algorithms such as alignment and phylogenetic reconstruction software (Tomii and Kanehisa, 1996).

Design of the database

Choice of database strategy There are generally two types of database employed to store and organize data: one is the relational database and the other is the XML database. Relational database technology is the older of these, and is therefore more common in current life science databases (e.g. GenBank, see Benson *et al.*, 2006, PDB, see Berman *et al.*, 2000). As a mature technology, relational databases are associated with sophisticated query languages (SQL) and development software (e.g. Oracle or MS Access). However, they suffer from inflexibility, requiring that all types of data are predefined: once the database has been created, subsequent changes are difficult to achieve and considered poor programming practice as they can easily disrupt database stability.

The newer technology of XML is based on a data description language (XML) designed to facilitate cross-platform data exchange of complex, non-homogeneous data types using customized, self-explanatory tags so that both computers and human can understand the semantics. Thus, XML databases handle irregularity of data well and are highly suited to development where not only the number of data items, but also the properties of these data items, are likely to change and expand as time proceeds (e.g. see the new Gene Expression Omnibus database from the NCBI, Barrett *et al.*, 2005).

A major, ongoing characteristic of amino acid research concerns the diverse biological and chemical roles that each can play: it is likely that new roles and associated measurements will continue to emerge as this science grows. In this context, an explicit goal of our database is to create a foundational resource that can expand in depth and breadth over time, as new research highlights new biophysical properties or new molecules pertinent to amino acid research. Thus, the XML is a natural choice for our work. Since our initial database is relatively small (388 entries), the slow speed that can arrive as the cost of XML flexibility is unnoticeable: orders of magnitude more information would be needed to render the speed of XML problematic, and history suggests that improvements in computing speed are likely to out match any perceptible reductions in speed caused by significant database growth.

Data structures of the database Since one XML document in the database corresponds to one amino acid, it is easy to add a new document/new amino acid containing considerable information. Figure 1 shows a schematic overview of the data structure we use to illustrate the relationships among XML elements in an AA-QSPR Db XML file. We use the common name and Chemical Abstracts Service (CAS) registry number (SciFinder: http://www.cas.org/SCIFINDER/) to identify an amino acid and use Simplified Molecular Input Line Entry System (SMILES: see Weininger, 1988), a linear structural representation, and molecular formula to provide general structure information. An XML tag created explicitly for this project ('foundin'), acts as a major classifier, defining biosynthetic, coded, abiotic and engineered amino acids by the source(s) from which an amino acid has been identified. For example, the standard amino acid alanine has been found in Murchison meteorite (Cronin and Moore, 1971) and prebiotic chemistry experiments (Miller, 1953), which suggest it can be synthesized abiotically; therefore, we label it as 'abiotic' and 'biosynthetic' inside 'foundin' tags (see Supplementary XML file of amino acid alanine). In the element of 'GeneralInfo', an attribute called 'coded' is used to differentiate whether the amino acid is one of the twenty standard amino acids. The biophysical properties of an amino acid are described in the XML element 'descriptors', which includes biophysical property name (e.g. $\log P$), associated values, whether these values are predicted or experimentally determined and a reference or source for the property value (those used here have been previously evaluated for accuracy, in Lu and Freeland, 2006b, showing $>95\%$ correlation for size and charge with experimentally determined equivalents).

In Supplementary material, we include the XML schema for the AA-QSPR Db. A primary aim of our database is to encourage community development: in other words, to encourage all interested parties to contribute new molecules and new information for existing molecules. We therefore provide simple web forms in which members of the user community can submit new information. To prevent vandalism, only registered users can upload new XML files into the database. However, the registration process is simple. After

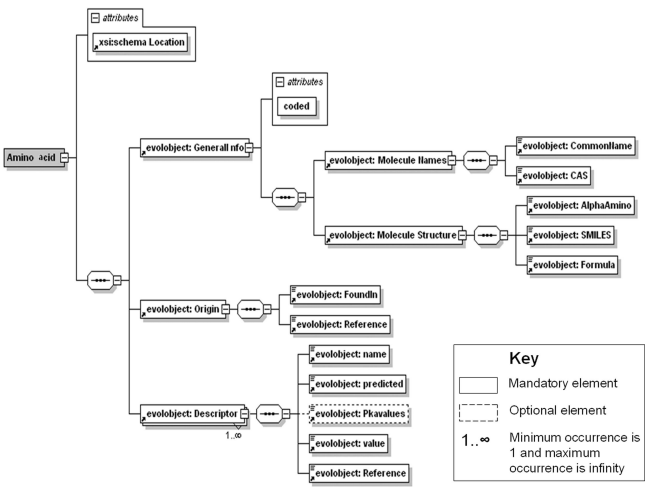

Fig. 1. Graphical representation of AA-QSPR data structure (generated by Altova XMLSpy software for data structure representation).

entering name, email addresses and affiliations, users will receive an email with a randomly generated password, which can be changed later. A correct match of username and password is necessary for a user to upload his/her data. The user-contributed amino acid XML files will contain the username of the contributors for future references. As a further safeguard, we back up the database once every 24 h, allowing us to restore it to a recent version if any major problems occur.

Database web interfaces and visualization/analysis tools Our database includes user-friendly web functions developed to help non-computer scientists navigate the website and database. An online help manual, which includes a comprehensive tutorial, is readily accessible to users. Clearly marked on the main page (http://www.evolvingcode.net:8080/AA-QSPR/html/), links lead users to the following functions: viewing XML entries in both XML and HTML formats; searching the database with keywords (e.g. amino acid common name); downloading data (including individual XML entries or entire data sets of the database) in ASCII format or as an XML schema; creating an XML file using an online form and uploading newly created XML files to the database (registration required); converting SMILES to Structures Data File (SDF) (Dalby *et al.*, 1992), on-the-fly, so as to view two-dimensional amino acid molecular structures with JMOL (http://jmol.sourceforge.net/) or for downloading to users' local machine. Furthermore, we implemented an online calculation of van der Waals volume and a connection to ALOGPS (Tetko *et al.*, 2005) Web Service for users to predict log *P* values. With the help of these web functions, our database serves as both a resource and a research platform that can enrich the knowledge base of the whole scientific community.

Our database is currently equipped with two visualization tools that help users investigate the relationships between the amino acids of their interest. One is an implementation of the 'KING' interactive three-dimensional vector graphics software (Davis *et al.*, 2004). This allows users to produce and manipulate interactive three-dimensional plots of 'chemical space' for user-defined sub-sets of amino acids according to any combination of van der Waals volume, pI and log *P*. The second visualization method is an incorporation of the open source-package 'TouchGraph' (http://www.touchgraph.com) which builds a minimum spanning tree from amino acids selected by users. This powerful analysis and visualization method has been widely used in many research areas to help researchers gain an intuitive insight about the complex relationships of interest (Tomii and Kanehisa, 1996; Knight *et al.*, 2006; see Bulka *et al.*, 2006 for a more detailed description of the method). Essentially, a user-defined, quantitative measure of the distance between amino acids is used to connect them into a tree structure in which adjoining elements are most similar to one another. Thus, for example, Fig. 2 shows a wide distribution of the 20 proteinaceous amino acids on major branches of the minimum spanning tree that is made up of 102 naturally occurring amino acids.

Both visualization tools were implemented in a way that can cope with the growing database, allowing users full control of which amino acids and which properties are incorporated into plots. Both facilitate easy interpretation of a visualization by allowing the user to define color-coded sets of amino acids for display (e.g. to contrast biologically

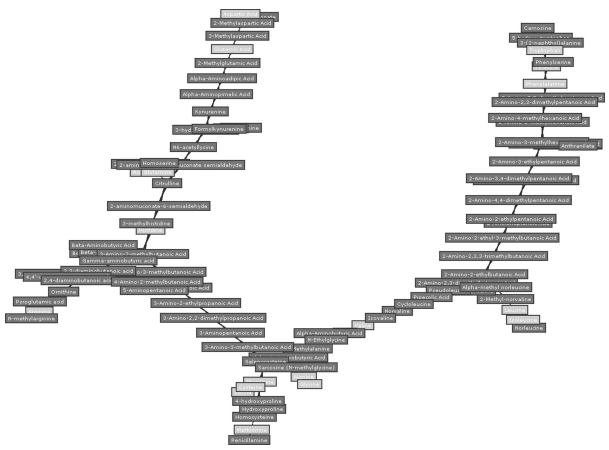

Fig. 2. Minimum spanning tree of 102 amino acids in AA-QSPR. Biologically encoded amino acids are shown in light gray boxes (this would be a user-defined color on the web).

coded amino acids within the super-set of those that are prebiotically plausible). Both tools further allow users to add, on-the-fly, any other molecules of specific interest so as to investigate the relationships between their own compounds and the amino acids of the database. Detailed instructions and examples of using these tools are available in online tutorial of AA-QSPR Db.

Example analyses

To illustrate the types of exploration that our database can support, here we present two simple, quick QSAR studies of peptides that include non-standard amino acids. Each recaptures the information reported from a more costly and laborious empirical study (Ufkes *et al.*, 1982; Asao *et al.*, 1987). One demonstrates the utility of the three fundamental biophysical properties that our database focuses on and the other demonstrates the speed and ease with which meaningful bioactivity results can be explored.

In a previous study (Ufkes *et al.*, 1978, 1982), 40 pentapeptides (Supplementary Table 1), including 10 that contain non-standard amino acids at one or two of five positions, were experimentally tested for their ability to potentiate bradykinin (a pharmacologically and physiologically active nine-mer peptide from the kinin group of proteins). Using the predicted amino acid size, charge and hydrophobicity values in our database, we created a matrix of 40 rows (one for each peptide) and 15 columns (for the three property values at each of five positions: see supplemental data). Using these as predictor (independent) variables, and the experimentally determined log RAI of each peptide (the logarithm of a relative potentiating activity index: Ufkes *et al.*, 1978, 1982) as the dependent variables, we then performed partial least squares (PLS) regression analysis on the data set. This approach is well established in chemistry for multivariate linear regression (Hattotuwagama *et al.*, 2006; Put *et al.*, 2006), and although there exist several variants of the precise method (e.g. one alternative to the PLS that we use is SIMPLS), these variations are equivalent if the response is uni-dimensional (Boulesteix and Strimmer, 2007). The PLS (performed by SAS 9.0) gave two clearly significant PLS components that together explained 78.4% of

the variance (65.2% and 13.2%, respectively) in the potentiating activity that previous empirical studies had reported. The weights that PLS assigned to the 15 predictors for the first extracted factor are in Supplementary Table 2. The correlation coefficient between PLS-predicted and experimentally measured peptide log RAI values is 0.89 (Fig. 3 and Supplementary Table 1): in other words, the database allowed us, as users, to recapture in seconds the findings of a laborious and expensive empirical study with 89% accuracy.

Next, we performed PLS on another data set of 48 dipeptides (Supplementary Table 3) for which a quantitative threshold of 'bitterness of taste' had been determined (Asao *et al.*, 1987). This time PLS generated a one-dimensional model that explained 82.9% of the experimentally reported variance in bitterness. For hydrophobicity (i.e. $\log P$), the two weights of its corresponding predictors were 0.41 and 0.56; for size (i.e. van der Waals volume), they were 0.43 and 0.57; for charge (i.e. pI), 0.02 and 0.17. The correlation coefficient between PLS-predicted and experimentally measured bitterness values is 0.91 (Fig. 4 and Supplementary Table 3).

Since there exist many computational chemistry programs for predicting biophysical properties (especially $\log P$), we repeated each of these analyses using the software which we previously found to have the second best accuracy (Lu and Freeland, 2006b) to gain some idea of the sensitivity of our data to error. Specifically, then, we replaced the $\log P$ values predicted by KowWin (Meylan and Howard, 1995), as found in our database, with those estimated by ALOGPS (Tetko *et al.*, 2005), which are not stored in our database. When using ALOGPS values, the correlation coefficients between PLS-predicted and experimental values were 0.82 for log RAI data set and 0.91 for bitterness. Thus, PLS predictions may vary slightly when they are performed on property values estimated by different computational programs (see Lu and Freeland, 2006b for previous investigation on various computational chemistry programs), though the underlying result does not.

Overall, these analyses are especially significant for their simplicity when placed against the increasing diversity of molecular descriptors that are entering QSAR studies (Jonsson *et al.*, 1989; Mei *et al.*, 2005). Thus, these two tests illustrate how our database can, in a few minutes, generate strong predictions of the results of relatively complex, expensive and time-consuming experiments.

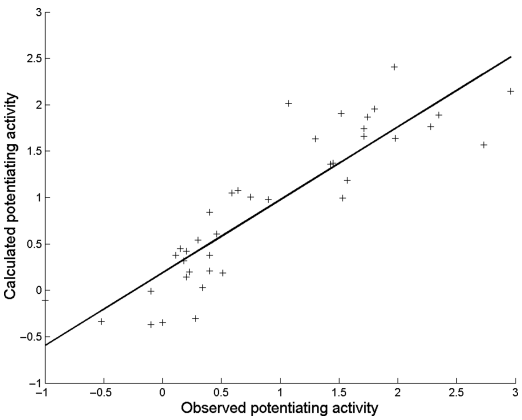

Fig. 3. Scatter plot of the observed and predicted potentiating activity of 40 pentapeptides, 10 of which include non-standard amino acids .The amino acid sequences of these peptides are provided in Supplementary Table 2. The correlation coefficient between our PLS-based predicted results and those of the previous experimentally determined values (Ufkes *et al.*, 1978, 1982) is 0.89.

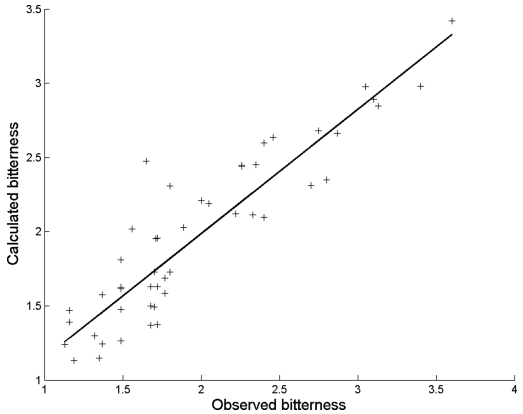

Fig. 4. Scatter plot of the observed and predicted bitterness of 48 dipeptides. The amino acid sequences of these peptides are in Supplementary Table 3. The correlation coefficient between our PLS-based predicted results and those of the previous experimentally determined values (Asao *et al.*, 1987) is 0.91.

Discussion

The AA-QSPR Db is a research tool designed to facilitate quick and easy exploration of amino acid 'chemical space' using modern web technology. Our aim is not, of course, to replace empirical studies; rather it is to offer rapid, safe and cheap explorations of amino acid chemical space which may act to focus the time and money associated with more detailed, empirical studies. Specifically, researchers anywhere in the world can now 'point and click' to rapidly select and explore various biophysical properties of any collection of amino acids so as to guide their analyses and experiments in protein design, origin-of-life research or bioinformatics.

Acknowledgements

We would like to thank Dr Michael New at NASA, Dr Gregurick at UMBC and Dr Boulesteix at Sylvia Lawry Centre for Multiple Sclerosis Research for their insightful input. This work is supported in part by grant NNG04GJ72G from Astrobiology: Exobiology and Evolutionary Biology.

References

Asao,M., Iwamura,H., Akamatsu,M. and Fujita,T. (1987) *J. Med. Chem.*, **30**, 1873–1879.

Barrett,T., Suzek,T.O., Troup,D.B., Wilhite,S.E., Ngau,W.C., Ledoux,P., Rudnev,D., Lash,A.E., Fujibuchi,W. and Edgar,R. (2005) *Nucleic Acids Res.*, **33**, D562–D566.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) *Nucleic Acids Res.*, **34**, D16–D20.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Biro,J.C. (2006) *Theor. Biol. Med. Model*, **3**, 15.

Boulesteix,A.L. and Strimmer,K. (2007) *Brief Bioinform.*, **8**, 32–44.

Bulka,B., desJardins,M. and Freeland,S.J. (2006) *BMC Bioinformatics*, **7**, 329.

Cronin,J.R., Gandy,W.E. and Pizzarello,S. (1981) *J. Mol. Evol.*, **17**, 265–272.

Cronin,J.R. and Moore,C.B. (1971) *Science*, **172**, 1327–1329.

Cronin,J.R., Pizzarello,S. and Yuen,G.U. (1985) *Geochim. Cosmochim. Acta*, **49**, 2259–2265.
Cronin,J.R. and Pizzarello,S. (1986) *Geochim. Cosmochim. Acta*, **50**, 2419–2427.
Cronin,J.R. and Pizzarello,S. (1997) *Science*, **275**, 951–955.
Czajgucki,Z., Andruszkiewicz,R. and Kamysz,W. (2006) *J. Pept. Sci.*, **12**, 653–662.
Dalby,A., Nourse,J.G., Hounshell,W.D., Gushurst,A.K.I., Grier,D.L., Leland,B.A. and Laufer,J. (1992) *J. Chem. Inf. Comput. Sci.*, **32**, 244–255.
Davis,I.W., Murray,L.W., Richardson,J.S. and Richardson,D.C. (2004) *Nucleic Acids Res.*, **32**, W615–W619.
Garrett,R.H. and Grisham,C.M. (1999) *Biochemistry*. Saunders College Publishing, Orlando, Florida.
Glavin,D.P. and Bada,J.L. (2001) *Astrobiology*, **1**, 259–269.
Grantham,R. (1974) *Science*, **185**, 862–864.
Guan,P., Doytchinova,I.A., Walshe,V.A., Borrow,P. and Flower,D.R. (2005) *J. Med. Chem.*, **48,** 7418–7425.
Haidacher,D., Vailaya,A. and Horvath,C. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 2290–2295.
Hattotuwagama,C.K., Toseland,C.P., Guan,P., Taylor,D.J., Hemsley,S.L., Doytchinova,I.A. and Flower,D.R. (2006) *J. Chem. Inf. Model* **46,** 1491–1502.
Hendrickson,T.L., de Crecy-Lagard,V. and Schimmel,P. (2004) *Annu. Rev. Biochem.* **73**, 147–176.
Jonsson,J., Eriksson,L., Hellberg,S., Sjostrom,M. and Wold,S. (1989) *Quant. Struct. Act. Relat.*, **8**, 204–209.
Kawashima,S., Ogata,H. and Kanehisa,M. (1999) *Nucleic Acids Res.*, **27,** 368–369.
Knight,C.G., Zitzmann,N., Prabhakar,S., Antrobus,R., Dwek,R., Hebestreit,H. and Rainey,P.B. (2006) *Nat. Genet.*, **38**, 1015–1022.
Link,A.J., Mock,M.L. and Tirrell,D.A. (2003) *Curr. Opin. Biotechnol.*, **14**, 603–609.
Lu,Y. and Freeland,S.J. (2006a) *Genome Biol.*, **7**, 102.
Lu,Y. and Freeland,S.J. (2006b) *Astrobiology*, **6**, 606–624.
Mei,H., Liao,Z.H., Zhou,Y. and Li,S.Z. (2005) *Biopolymers*, **80**, 775–786.
Meierhenrich,U.J., Munoz Caro,G.M., Bredehoft,J.H., Jessberger,E.K. and Thiemann,W.H. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 9182–9186.
Meylan,W.M. and Howard,P.H. (1995) *J. Pharm. Sci.*, **84**, 83–92.
Miller,S.L. (1953) *Science*, **117**, 528–529.
Miller,S.L. (1986) *Chem. Scr.*, **26B**, 5–11.
Put,R., Daszykowski,M., Baczek,T. and Vander Heyden,Y. (2006) *J. Proteome Res.*, **5,** 1618–1625.
Qiu,J.X., Petersson,E.J., Matthews,E.E. and Schepartz,A. (2006) *J. Am. Chem. Soc.*, **128**, 11338–11339.
Summerer,D., Chen,S., Wu,N., Deiters,A., Chin,J.W. and Schultz,P.G. (2006) *Proc. Natl Acad. Sci. USA*, **103**, 9785–9789.
Tetko,I.V. *et al*. (2005) *J. Comput. Aided Mol. Des.*, **19**, 453–463.
Tomii,K. and Kanehisa,M. (1996) *Protein Eng.*, **9**, 27–36.
Ufkes,J.G., Visser,B.J., Heuver,G. and Van der Meer,C. (1978) *Eur. J. Pharmacol.*, **50,** 119–122.
Ufkes,J.G., Visser,B.J., Heuver,G., Wynne,H.J. and Van der Meer,C. (1982) *Eur. J. Pharmacol.*, **79**: 155–158.
Uy,R. and Wold,F. (1977) *Science*, **198**, 890–896.
Venton,B.J., Robinson,T.E., Kennedy,R.T. and Maren,S. (2006) *Eur. J. Neurosci.*, **23,** 3391–3398.
Weber,A.L. and Miller,S.L. (1981) *J. Mol. Evol.*, **17**, 273–284.
Wang,L., Xie,J. and Schultz,P.G. (2006) *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 225–249.
Weininger,D. (1988) *J. Chem. Inf. Comput. Sci.*, **28**, 31.

Received January 1, 2007; revised April 9, 2007; accepted May 17, 2007

Edited by Philipp Holliger