SYSTEMATIC NAMES FOR SYSTEMS BIOLOGY

RICHARD CAMMACK

Department of Life Sciences, King's College London, UK
E-Mail: richard.cammack@kcl.ac.uk

Received: 14th April 2004 / Published: 1st October 2004

ABSTRACT

The aim of systematic nomenclature is to provide a name for each entity, such as a metabolite, an enzyme, or a measured quantity. There are different requirements for biochemical nomenclature, depending on how the name or symbol is to be stored and communicated, by written, printed, or spoken word, as a diagram, or as computer-readable data. Names are often related to biological function, structure or evolutionary relationships; nomenclature follows classification. For interaction with computers and databases, identifiers should be searchable, and referred to an authoritative source. The requirements for nomenclature are distinct from those of a dictionary, where the criterion for inclusion of a word is that it is used. When proposing systematic nomenclature, timely intervention is important, and much effort should be devoted to ensuring acceptance of within the scientific community.

INTRODUCTION: THE NEED FOR NOMENCLATURE SYSTEMS

Communication in science relies on having consistent and recognizable terminology, units and symbols. This is particularly important in multidisciplinary areas such as systems biology. For mathematicians, biochemists, bioinformaticists, chemists and other scientists to communicate, we need a standard and unambiguous name for each entity or concept. Inevitably, new species, compounds and concepts will be given different names at first, by research workers in different subjects. There needs to be agreement as to what these names should be, and there needs to be a mechanism to connect these "preferred" terms to other terms found in the literature. This is an important function of databases, glossaries and dictionaries. Interoperability of databases also depends on consistent nomenclature. In other words, we need to know we are talking about the same thing.

Published in „Experimental Standard Conditions of Enzyme Characterizations“, M.G. Hicks & C. Kettner (Eds.), Proceedings of the 1st Int'l Beilstein Symposium on ESCEC, Oct. 5th - 8th 2003, Rüdesheim, Germany

There will be a large number of enzymes, proteins and other cell factors to be named in the future. In the genome of the well-understood organism *Escherichia coli*, there are about 36% of the predicted gene products for which the function is still completely unknown. Moreover some of the present annotations will probably prove to be inaccurate, and there are many new species that have hardly been explored. So we need a system in place to agree on good names for these entities.

Table 1. Criteria for a good name or descriptive phrase.

Essential	Advisable
unique	good search term
infinitely extendable	memorable
not be easily confused with anything else	reflects structure or function
open source	decided by international authority
not copyright or a trademark	not obscene or hilarious in any language

There are many different requirements for a system of nomenclature, and no system is perfect. However for the purposes of this discussion it is useful to list the desirable characteristics of an effective nomenclature system (Table 1).

WHAT MAKES A GOOD NAME?

There should be standards for nomenclature and symbols in systems biology. A good name for an entity or phenomenon can crystallize our thoughts about it.

What makes a good name ? This depends on:

- the medium in which it is to be presented (Table 2)
- recommended or informal nomenclature
- who is intended to use it:

laboratory specialists

specialist community

wider biochemical community

scientific community

Table 2. Different formats, and nomenclature issues.

Format	Problem	Example
Printed word	Character formats	l and 1, O and 0
Handwritten (lab notebook)	Legibility of symbols	*v* and *V* (kinetics)
Spoken word	Pronounceable	Sulfenate/sulfinate
Diagrams	Computer readability	Metabolic maps
Structures	Standard representation	
ASCII text, internet	Special characters	Greek letters, italics, subscripts etc.
Database	Generally only use ASCII	Consistent use, e.g. Unicode
Proprietary, Trademarks	Interoperability	Adrenaline/epinephrine

Some names can be helpful, because they invoke an analogy with another well-known term. "Polymerase chain reaction" (PCR) is such a name, as it suggests a similarity to other chain reactions of chemistry and nuclear physics: one molecule leads to more and more products. On the other hand some names can trip up the uninitiated. "Real-time" is a well-known term in computer science to describe a process that a computer monitors *as it occurs*. Someone reading about "real-time PCR" might expect it to be some sort of continuous or instantaneous measurement. In fact the most significant feature of real-time PCR is that it provides a *quantitative* measure of the amount of DNA. Add to this the frequent use of "RT-PCR" for both this method and "reverse-transcription PCR", and there is scope to confuse the uninitiated.

Names originating from laboratory jargon often cause problems.

- An example of informal notation is the letter "p" followed by a number. Often the number represents the apparent molecular mass, in kilodaltons, on SDS-polyacrylamide gel electrophoresis. An example is the intensively studied tumour-suppressor gene p53 (actually a tetramer, so its molecular mass is 212 kDa). Such names are not extendable, although a series of related proteins, p63 and p73 have been described. Sometimes the "P" (in capitals) stands for "pigment". The well-known cytochrome P450 is named for the Soret peak in its carbon monoxide difference spectrum. Being an enzyme and not just an electron-transfer protein, it is not recognized as a cytochrome in systematic nomenclature [1].
- X+number, where X stands for species: "H" could be horse, horseradish or Hansenula. There are simply too few letters in the alphabet!

- TLA's (three-letter abbreviations) and similar short names are a source of confusion. Such abbreviations have a large number of hits when searched in PubMed, but this may conceal a range of different meanings.
  Since there are only 17,576 permutations, the chances of having more than one meaning is high (Table 3). Often the meaning of the abbreviation is buried in the text of a paper, which makes it difficult for the reader to find.

Table 3. An example of ambiguity with TLA's.

ACF	ATP-utilizing chromatin assembly and remodeling factor
	APOBEC-1 complementation factor
	aberrant crypt foci of the colon
	anticoagulation factor
	2-[(2-amino-4-chloro-5-fluorophenyl)thio]-N,N-dimethyl-benzenmethanamine
	anterior corpectomy with fusion
	N-acetyl phenylalanine
	accessory colonization factor

Acronyms are useful if they are good search terms. They should preferably not be the same as common words, e.g. WAVE designed for DNA fragment analysis. This helps as a mnemonic, but makes them difficult to find in literature searches.

Names with complex syntax, such as capitals and small letters, mixed with numerals, have the problem that they are easy to forget, and mistakes are often made. One only has to think of the complex passwords that are required to log onto some secure data systems ("Forgotten your password again ? Click here..."). But biochemical nomenclature which has no systematic basis is also difficult to use consistently. An example shown in Table 4 is the proton-translocating ATP synthase of mitochondria, known as the F_oF_1 ATPase (class EC 3.6.3.14). This name dates from a time when manuscripts were typewritten, and there were inconsistent uses of characters, such as capital "O" or zero. In the original form the lower case subscript "o" stood for oligomycin-sensitive, and F_1 represented the large water-soluble part of the protein complex. Thus, many different variants have appeared in the literature.

Table 4. Synonyms in the literature for the proton-translocating ATPase.

	Frequency of use	
	Web of Science®*	Google™*
F_1 ATPase	531	7590
F-1 ATPase	508	4520
F0F_1 ATPase	225	911
FoF_1 ATPase	30	426
F_1Fo ATPase	30	554
F0/F_1 ATPase	3	167
F0F_1	266	1750
FOF_1	51	4680
F_1F0	279	3610
F_1Fo	40	772
ATP synthase	1365	63600
ATP synthetase	22	31500

This raises a number of points about the differences in usage of names in the written word and in computers. A human reader will easily recognize that all the terms in Table 4 probably represent the same thing. However character-matching software obviously regards them as distinct, since the hits in Table 4 were found by searching computer databases. In order for a database of enzymes to provide an accurate representation of the literature on the subject, it must include all the variant forms. In fact the number of variants is actually much greater than indicated in Table 4, because the search engines used to determine the number of "hits" take no account of upper and lower case, italics, subscripts and superscripts, greek letters and other symbols.

The requirements for a distinct written name, and a good search term, mean that compromises are being made in terminology. Databases may employ some form of encoding to distinguish variants in syntax, although there is no consistent practice. In the literature, features of punctuation, such as italics in species names and foreign phrases, are increasingly being omitted. A recent such recommendation is that the italics representing the source organism in symbols for restriction enzymes should be omitted for example *Eco*R1 would be EcoR1 [2].

SYSTEMATIC CHEMICAL NAMES

In chemistry, the most important characteristic of a compound for classification purposes is usually its structure. Chemical compounds were first given arbitrary names, as they were identified, but the number of these had become unsustainable during the 19th century. International efforts to create an acceptable system of nomenclature of organic compounds date back at least as far as the Geneva Convention of 1892, and have been extended and refined ever since [3]. This was followed by recommendations for inorganic, physical, organometallic and macromolecular chemistry. These systems of chemical nomenclature are principally aimed at providing a name, which can be written or spoken, that defines every compound. The names of compounds defined by the IUPAC systems have legal standing, for example in patents.

A useful introduction to the principles of chemical nomenclature is provided by the Guide to IUPAC recommendations [4]. Usually the name of a compound is derived from a parent compound, with substituents at positions defined by a numbering system. The formalisms are continually being reviewed and extended, to describe new classes of molecules such as fullerenes.

Computer databases are now an indispensable part of sciences such as organic chemistry, where enormous numbers of new compounds are synthesized. They allow information on structures, spectroscopic and other physical properties, to be assembled in an accessible way. The new areas of science such as systems biology would not be possible without computer databases. Databases such as the CAS (Chemical Abstracts Service) Registry (http://www.cas.org/EO/regsys.html) list all compounds, including biochemical compounds and gene sequences. The CAS number is an identifier, and can only be understood in the context of the database. It provides a means of cross-referencing different names for a compound. A biochemical compound such as glucose may have several CAS Registry numbers, reflecting the different enantiomers, open-chain and ring structures that interconvert spontaneously in solution.

Generally, chemists who are non-experts in nomenclature find it easier to visualize a chemical structure than to interpret a systematic chemical name. It is easy to make mistakes when deriving a chemical name. Changing a bond in a ring structure, for example, can completely change the numbering of the rest. Increasingly the task of converting structures to names, and names to structures, is being taken over by software, such as the programs used for drawing chemical structures, which implement the rules of chemical nomenclature.

As an alternative to systematic chemical names, linear notations for molecular structures have been developed. SMILES^{TM} strings (http://www.daylight.com/) are a well-established notation for chemical structures. Based on a set of simple rules, they are readily generated for a particular molecule. This can also be done by proprietary software packages.

If the data on chemical compounds is to be stored on databases, it becomes less important that the name used is readable by humans. A recent innovation is the ICHI (IUPAC Chemical Identifier) or INCHI (IUPAC-NIST Chemical Identifier) [5,6]. This is an ASCII string, generated by a computer algorithm, that uniquely defines a compound. In contrast to the SMILES system, where often several valid strings can be written for a compound, every chemical structure yields a unique INCHI. The INCHI is open-source, whereas the software to create SMILES strings is proprietary, and even some of the strings themselves are copyright. The INCHI has the status of a IUPAC project at the moment, and software to use it has yet to be developed. However this identifier, if adopted widely, should be extremely useful in databases.

An important feature of the description of an entity in a database is its *identifier*. This is an invariant label for the entity within the data system. It should be extensible, that it has sufficient letters and digits to encompass all examples that could possibly be encountered. It is important to recognize that an identifier should be devoid of any other information. Numbers used as identifiers are never re-used within the database. If the entity is given another name, it can be traced back through the system. Often there are several ways of naming a single compound. These should all be linked to the same identifier. Databases also use preferred names for compounds, a feature known as *controlled vocabulary*.

SYSTEMATIC NAMES IN BIOCHEMISTRY AND MOLECULAR BIOLOGY

In bioinformatics, systematic chemical names would be unwieldy and non-intuitive. Moreover the experimental data on the structures and characteristics of biological molecules, is disparate, incompletely defined, and distributed among different online databases. When working with information on the Internet, interoperability is a watchword. This means that, while working on a database, information in other databases should be only a few clicks away. The need for any conversion software or password access slows the process down enormously [7].

Often the same or similar molecule is given a different name, for example if derived from a different species. Computer databases can manage this complexity by storing and manipulating lists of synonyms, as part of their controlled vocabulary.

Systematic, functional nomenclature implies a classification. A classification of a gene product may be on the basis of function, molecular structure, phylogeny or genes. There should be a hierarchy of such criteria, otherwise conflicts will arise where one criterion implies that an entity belongs in one class, and another criterion would put it in another. Genes often have a multiplicity of names. More than one name is used for similar gene products in different species, or even from the same species. Organizations such as the HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/) aim to simplify this multiplicity as much as possible. They publish, and apply, guidelines for nomenclature, in parallel with those for the mouse and other genomes [8].

Biochemical nomenclature has been an ongoing activity since the 1950s. Its original purpose was to arrive at a more consistent terminology in the literature and the textbooks. It is recommended in the instructions to authors of biochemical journals. Editors have a part to play in encouraging authors to use recommended nomenclature and terminology. Currently these activities are coordinated by the Nomenclature committee of IUBMB (NC-IUBMB) and the Joint Commission on Biochemical Nomenclature (JCBN) [9]. The two committees work together, to set up panels for specific nomenclature. This has led to the publication of reports on the nomenclature of proteins, carbohydrates, nucleic acids and other compounds, published in book form [10] and more recently on the web (http://www.chem.qmul.ac.uk/iubmb/). Newsletters are posted on the website (http://www.chem.qmul.ac.uk/iubmb/newsletter/).

The EC list of enzymes, one of the activities of the committees, provides a good example of a functional classification and system of nomenclature (http://www.chem.qmul.ac.uk/iubmb/enzyme/). It is described in the article by Sinead Boyce et al. in this book. The basis of classification is the reaction catalysed. An entry in the EC list denotes simply that an enzyme has been shown to exist, that catalyses the approach to equilibrium of a specific biochemical reaction. The EC number identifies the reaction classified. The EC number lends itself naturally to computing, and there are databases that use it as the primary method of searching, e.g. INTENZ, part of the Expasy database of protein structure and function (http://www.ebi.ac.uk/intenz/).

The EC class of an enzyme is arrived at by application of a set of rules [11]. Other secondary criteria such as cofactor composition have occasionally been invoked to distinguish between enzymes, but in most cases they are not admitted since they may cause confusion. The EC list is a classification of enzymes that are demonstrated to exist, rather than a list of possible reactions. Because the EC class may change in the light of new knowledge about the enzyme, it is therefore not an identifier, for database purposes.

The nomenclature committees operate interactively with the biochemical community. The process of classification often begins with the submission of information from someone who is working on that enzyme. There is an exchange of information that leads to the checking of the enzyme details, creation and correction of a draft entry. To fulfil the requirements for public consultation, proposed entries or revisions of the enzyme list are displayed on the website at www.chem.qmul.ac.uk/iubmb/enzyme/newenz for two months, while biochemists (including those who proposed the entries) are invited to comment. After the consultation period the entry is corrected and put into the enzyme list. (Fig. 1). EC numbers are never re-used even if they are finally not approved or they become superseded. The progress of any changes to EC numbers is traceable through information on the website.

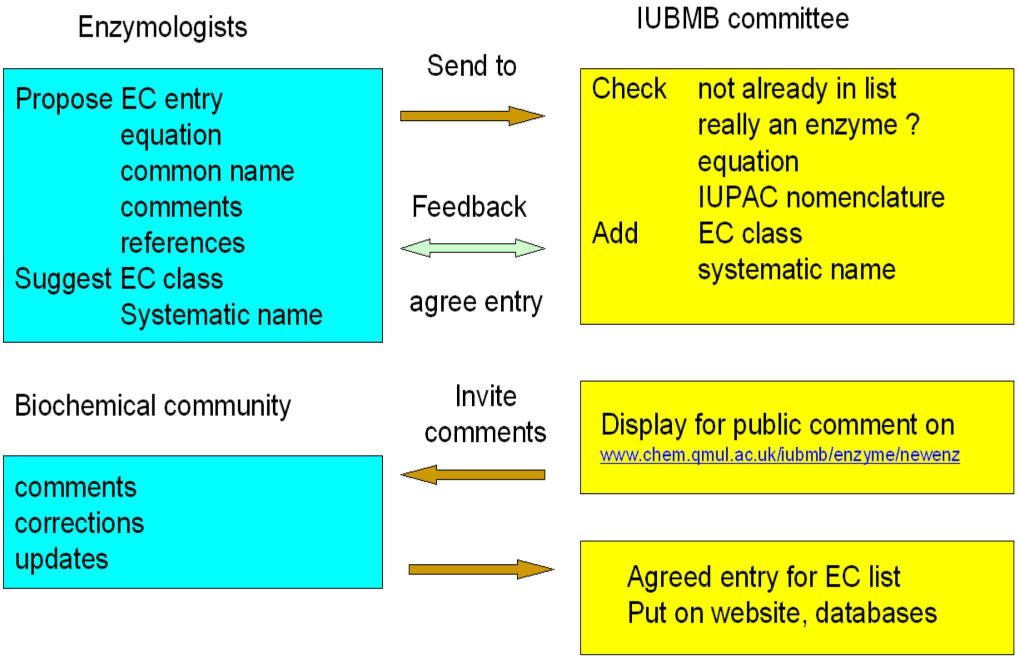

Figure 1. Flow chart for classification of an enzyme

DICTIONARIES AND GLOSSARIES

Along with the development of online bioinformatics databases, there is still a need for dictionaries and glossaries. Although an unfamiliar word can be found by using search engines such as Google and the literature databases, a dictionary describes how it is normally used [12]. The inclusion of a misleading or ambiguous term provides the opportunity for cross-references to alternative or recommended names.

The principal criterion for inclusion of a new biochemical term in a dictionary or glossary is that it is widely used in the laboratory and in the literature. Literature searches provide a means to check the frequency with which terms are used, by their prevalence in the titles, keywords and abstracts of the relevant journals in biochemistry and molecular biology. For the second edition of the *Oxford Dictionary of Biochemistry and Molecular Biology*, the informal criterion being applied is that, for inclusion, a neologism should be mentioned at least 10 articles per year in titles, keywords and abstracts of the relevant journals.

ACCEPTANCE

The development and acceptance of nomenclature standards has been a gradual process. It is human nature to be reluctant to abandon familiar names, albeit they are non-systematic or even misleading. The work of developing internationally agreed standards has been undertaken by international bodies, particularly IUPAC, which has a formal process of public review before they are accepted.

Finally, it is important to remember that setting a standard does not necessary lead to compliance. It may be that everyone is talking about the same thing, but not using preferred nomenclature. It is not unknown for official recommendations to have such low levels of acceptance that they are finally forgotten. Generally this occurs when the usage of other names, units and symbols has become established and the scientific community does not perceive a need for new names, units and symbols. However if the new terms are easier to explain, and more intuitive to understand, new generations of students will accept them. Timely intervention is important: not too early when the compounds are inadequately understood; not too late when misleading terms have become embedded in the literature and databases.

References

[1] Palmer, G. (1989) Nomenclature of electron-transfer proteins. *J. biol. Chem.* **267**:665-677.

[2] Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S.K., Dryden, D.T.F., Dybvig, K., et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucl. Acids Res.* **31**:1805-1812.

[3] Panico, R., Richer, J.C., Powell, W.H. (1993) *A Guide to IUPAC Nomenclature of Organic Compounds*. Blackwell, Oxford.

[4] Leigh, G.J., Favre, H.A., Metanomski, W.V. (1998) *Principles of Chemical Nomenclature: a Guide to IUPAC Recommendations*. Blackwell, Oxford.

[5] Adam, D. (2002) Chemists synthesize a single naming system. *Nature* **417**:369.

[6] Stein, S.E., Heller, S.R., Tchekhovskoi, D.V. (2001) Toward the development of a standard chemical identifier. *Abstracts of Papers of the A.C.S.* **222**:5.

[7] Berendsen, H.J.C. (2003) Inter-union bioinformatics group report. *Acta Crystallogr. Section D-Biol. Crystallogr.* **59**:777-782.

[8] Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W., Povey, S. (2002) Guidelines for human gene nomenclature. *Genomics* **79**:464-470.

[9] Cammack, R. (2000) The biochemical nomenclature committees. *IUBMB Life* **50**:159-161.

[10] Liebecq, C. (1992) *Biochemical Nomenclature and related documents*. Portland Press, London.

[11] Webb, E.C. (1992) *Enzyme Nomenclature*. Academic Press, San Diego.

[12] Smith, A.D., Datta, S.P., Howard Smith, G., Campbell, P.N., Bentley, R., McKenzie, H.A. (eds) (1997) *Oxford Dictionary of Biochemistry and Molecular Biology.* Oxford University Press, Oxford.