ISSN 1367-4803

BIOINFORMATICS

VOLUME 16 NUMBER 1 JANUARY 2000

MAY

SPECIAL HISTORY ISSUE

OXFORD
UNIVERSITY PRESS

Uncover a hidden world

Like sifting sand for a jewel, research science can be a lifetime of bench work for one moment's flash of discovery. Your lifetime has just been extended... Welcome to the exciting world of research on the web. LabOnWeb's new platform, **InstantRACE™**, is not about e-commerce or portals.
It's about your work. Found a new sequence?
Got positives from a chip experiment?
Exploring a gene you know very little about?

Go to www.labonweb.com and you will be able to:

- Elongate sequences to full-length genes, using IRACE unique sequence extension tool.
- Analyze expression patterns, based on SAGE technology, and optimize DNA chip design.
- Understand role and function through the Functional Analysis Robot.

Sometimes, just a little innovation uncovers big secrets...

LabOnWeb.com
The first Research Engine on the web

LabOnWeb .com

Powered by compugen

BIOINFORMATICS **CONTENTS** Volume 16 number 1 2000

EDITORIAL — The Editors — 1

ORIGINAL PAPERS

The early days of bioinformatics publishing — 2
R.J.Roberts

Earliest pages of bioinformatics — 5
E.N.Trifonov

A historical perspective on gene/protein functional assignment — 10
T.C.Hodgman

DNA binding sites: representation and discovery — 16
G.D.Stormo

On the trail of protein sequences — 24
R.F.Doolittle

History of the enzyme nomenclature system — 34
K.Tipton and S.Boyce

The early introduction of dynamic programming into computational biology — 41
D.Sankoff

Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! — 48
A.Bairoch

A brief history of genome research and bioinformatics in France — 65
A.Danchin

INSTRUCTIONS TO AUTHORS

© Oxford University Press 2000

BIOINFORMATICS

Vol. 16 no. 1 2000
Pages 34–40

History of the enzyme nomenclature system

Keith Tipton and Sinéad Boyce

Department of Biochemistry, Trinity College, Dublin 2, Ireland

Abstract

Naming things is essential for people to understand one another, no matter what language or field of interest is involved. This is as true for enzymes, genes and chemicals as it is for birds, food, flowers, etc. Effective communication requires a lack of ambiguity, but, in practice, ambiguities abound even between people who use the same language in different parts of the world, or even within the same country. Whereas ambiguities in the words used for common objects or actions have been the basis for many, more-or-less memorable jokes, they can also cause a great deal of confusion. Such linguistic chaos is welcomed by many as being a part of a diverse heritage that should be preserved at all costs to prevent us from descending into Orwellian 'newspeak'. However, in the sciences, there are distinct advantages in others being able to understand what one is doing. Many groups have stressed the need for standardized, universally accepted systems of nomenclature in chemistry, genetics, enzymology, etc. However, it is the universal acceptance that usually causes the problem. It is rare to find people who will admit that they find nomenclature to be an interesting subject, but many who profess contempt for it will get very excited if it is suggested that their pet nomenclature should be changed in the interest of clarity or uniformity. This account will consider the development of the enzyme nomenclature system, its benefits, shortcomings and future prospects.

Contact: *ktipton@tcd.ie*

Pre-history

In 1833, Payne and Persoz found that an alcohol precipitate of a malt extract contained a substance that converted starch into sugar. This was the first recognition of an enzyme and they named it diastase, a term that came to be used for enzymes in general. Payne and Persoz (1833) subsequently proposed that the names of enzymes take the last three letters of diastase, and that the suffix 'ase' should be appended to a root that indicated which substance was being acted upon by the enzyme. This method of naming enzymes has continued to the present day, except for proteolytic enzymes, which often end with 'in'. The term 'enzyme' is itself derived from the Greek for 'in yeast' and was coined in 1876 by W. F. Kühne. At that time, there were two schools of thought regarding enzymes. Liebig thought that fermentation enzymes were chemical substances found in yeast (unorganized ferments), whereas Pasteur believed that the process of fermentation could not be separated from the living cell (organized ferment). Kühne did not favour either hypothesis but was making the point that there was something in yeast that could exert an activity. The controversy continued until the fermentation system was isolated in a cell-free extract. The credit for this discovery has been the subject of one of those unseemly squabbles in which the ego is felt to be more important than the advance. von Manassein would appear to have priority as she had demonstrated this phenomenon and given an oral presentation on it in 1871. However, over a decade later, in 1897, Buchner claimed priority and dismissed von Manassein's claims, on the grounds that her sterile sugar solutions were undoubtedly contaminated with micrococci and, thus, did not constitute a cell-free system. At this stage of development, such an unedifying squabble may be of interest only to historians of science, but further details can be found in Cornish-Bowden (1997).

Early-days

The naming of enzymes was not systematic in its early days. Enzymes were given trivial names that often meant little or were ambiguous. Many different enzymes were given the same name and, conversely, several names were given to the same enzyme, leading to much confusion. In general, the suffix 'ase' was added to the name of the substrate, as in the case of urease (the first enzyme to be crystallized) or else the name gave some indication of the reaction catalysed, e.g. glucose oxidase. However, several other enzymes were given names that were decidedly unhelpful, for example, rhodanese (now called thiosulfate sulfurtransferase). Such names gave no indication of either the substrate involved or of the reaction taking place and the tendency to coin apparently meaningless names still persists, for example, barnase does not hydrolyse outhouses but is the extracellular ribonuclease secreted by a strain of *Bacillus amyloliquefaciens*.

Although such arcane names may help to keep the subject area esoteric, the rapid increase in the number of enzymes being discovered necessitated the development of some method for naming them systematically. In the

© Oxford University Press 2000

1950s, two groups of enzymologists set about addressing this problem. The first, instigated by Otto Hofmann-Ostenhof (1953), classified enzymes using a system based on the number of molecules involved in the reaction. He proposed the following three general classes:

1. *Hydrolases, transferases and oxidoreductases* (of the general reaction type A + B = C + D)
2. *Lyases and synthases* (A = B + C)
3. *Racemases* (A = B)

Malcolm Dixon and Edwin Webb, who were compiling a list of all known enzymes for their influential book *Enzymes*, noted that, despite the relatively large number of enzymes, the number of types of reaction involved was quite small. They also classified enzymes into three broad groups depending on the type of reaction catalysed: the hydrolysing, transferring and other enzymes. The latter group included the synthetases (enzymes catalysing synthetic reactions linked with the breakdown of ATP or GTP), stereoisomerases and enzymes that added groups to double bonds. The oxidizing enzymes were thought of as transferring hydrogen and were therefore categorized with the other transferring enzymes (Dixon and Webb, 1958). At the time of going to press, they had categorized 659 enzymes and, besides placing them in the above classes, they numbered them serially from 1 to 659.

(A) *Hydrolysing enzymes* (221 enzymes)

1. Enzymes hydrolysing –CO–NH– links (Numbers 1–69)
2. Enzymes hydrolysing amines, etc. (Numbers 70–88)
3. Enzymes hydrolysing ester links (Numbers 89–137)
   (a) *Carboxylic esters* (Numbers 89–105)
   (b) *Phosphoric esters* (Numbers 106–124)
   (c) *Sulphuric esters* (Numbers 125–128)
   (d) *Thiolesters* (Numbers 129–135)
   (e) *Phenolic esters* (Numbers 136 and 137)
4. Enzymes hydrolysing glycosidic links (Numbers 138–176)
5. Enzymes hydrolysing acid anhydride links (Numbers 177–186)
6. Enzymes catalysing non-oxidative decarboxylations (Numbers 187–213)
7. Enzymes hydrolysing other links (Numbers 214–221)

(B) *Transferring enzymes* (358 enzymes)

1. Enzymes transferring hydrogen (oxidoreductases) (Numbers 222–402)
   (a) *Enzymes oxidising –CHNH$_2$.COOH to –CO.COOH* (Numbers 222–230)
   (b) *Enzymes oxidising other nitrogenous groups* (Numbers 231–259)
   (c) *Enzymes oxidising –CHOH to –CO–* (Numbers 260–337)
   (d) *Enzymes oxidising)CH–CH$_2$– to)C=CH–* (Numbers 338–345)
   (e) *Enzymes oxidising or reducing sulphur groups* (Numbers 346–353)
   (f) *Other oxidising enzymes* (Numbers 354–402)
2. Enzymes transferring nitrogenous groups (Numbers 403–426)
   (a) *Enzymes transferring amino groups (transaminases)* (Numbers 403–420)
   (b) *Enzymes transferring amidine groups* (Number 421)
   (c) *Enzymes transferring amino acids, peptides, etc.* (Numbers 422–426)
3. Enzymes transferring phosphate groups (Numbers 427–503)
   (a) *Enzymes transferring phosphate groups from nucleosidetriphosphates (phosphokinases)* (Numbers 427–474)
   (b) *Enzymes transferring pyrophosphate from ATP (pyrophosphokinases)* (Numbers 475 and 476)
   (c) *Enzymes apparently catalysing intramolecular phosphate transfer (phosphomutases)* (Numbers 477–482)
   (d) *Enzymes transferring substituted orthophosphate groups of triphosphates, leaving inorganic pyrophosphate (pyrophosphorylases)* (Numbers 483–495)
   (e) *Enzymes catalysing other phosphate-transfer reactions* (Numbers 496–503)
4. Enzymes transferring acyl groups (transacylases) (Numbers 504–524)
5. Enzymes transferring glycosyl groups (transglycosylases) (Numbers 525–560)
6. Enzymes transferring coenzyme A (Numbers 561–563)
7. Enzymes transferring methyl groups (transmethylases) (Numbers 564–567)
8. Enzymes transferring other groups (Numbers 568–579)

(C) *Other enzymes* (80 enzymes)

Corteva Exhibit 2020 - Page 3 of 6
Inari v. Corteva - PGR2023-00022

1. Enzymes catalysing synthetic reactions linked with the breakdown of ATP or GTP (Numbers 580–603)
2. Enzymes adding groups to double bonds (Numbers 604–642)
3. Enzymes catalysing changes of steric configuration (stereoisomerases) (Numbers 643–655)
4. Enzymes catalysing other reactions (Numbers 656–659)

The list also provided information about the cofactors and prosthetic groups required by each enzyme, whether or not it had been crystallized, the organism or tissue-type where it had been found and some information about the substrate or reaction catalysed.

This was the beginning of the current enzyme classification and nomenclature system, with enzymes divided into groups and subgroups according to the nature of the reaction catalysed. However, the combination of such a division with a consecutive numbering system would not allow for any enzymes subsequently identified to be assigned a meaningful number in its class unless all those in subsequent classes were renumbered. For example, the next available number for a newly discovered peptidase might be 660 and so that class would then include Numbers 1–69 and 660.

Dixon and Webb recognized the artificiality of this division into categories, and pointed out that hydrolases could also be regarded as transferases, since they catalyse the transfer of one part of a molecule to water and, similarly, many of the reactions catalysed by the enzymes in their group (C) also involve transfer reactions. If one wanted to be pedantic, one could say that over 90% of the enzymes in that list could be regarded as being transferases. However, such a classification would not have been very useful and is not consistent with the way in which biochemists usually think about reactions.

The present format

In 1956, after being approached by Dixon and Hoffmann-Ostenhof, the President of the International Union of Biochemistry, Marcel Florkin, established an International Commission on Enzymes to tackle the problems of classification and nomenclature. The Commission's initial terms of reference were: 'To consider the classification and nomenclature of enzymes and coenzymes, their units of activity and standard methods of assay, together with the symbols used in the description of enzyme kinetics.'

It was an ambitious remit and not all these aspects have been brought to a satisfactory, or generally accepted, completion. A major role for this Commission was to devise a series of rules/recommendations that could be implemented for the systematic naming of old enzymes and according to which new enzymes could be classified. They considered several possible ways of categorizing enzymes, such as systems based on what was then known about the chemical nature of the enzyme (flavoprotein, haem protein, copper protein, etc.) or the chemical nature of the substrate (nucleotide, carbohydrate, protein, etc.) before opting for a system that was based on the *overall* reaction catalysed, which meant that intermediates formed during the reaction were not taken into consideration.

In 1958, they produced an interim report, which was approved and finalized in 1961 and published in the second edition of *Enzymes* (Dixon and Webb, 1964) as well as in a separate book covering enzyme nomenclature and the units and the symbols used in enzyme kinetics (Enzyme Nomenclature, 1965). In this, they extended the earlier system of Dixon and Webb (1958) by classifying enzymes into the following six categories:

Class	Name	Reaction catalysed
1.	Oxidoreductases	$AH_2 + B = A + BH_2$ or $AH_2 + B^+ = A + BH + H^+$
2.	Transferases	$AX + B = A + BX$
3.	Hydrolases	$A\text{–}B + H_2O = AH + BOH$
4.	Lyases	$A = B + X\text{–}Y = A\text{–}B$ $\quad\quad\quad

suggested by another committee (of which we were fortunately not members) was to work through each such discrepancy in turn and let different national groups alternate in choosing the names to be adopted. If this is in fact true it makes the reflective, but somewhat ponderous, operation of the JCBN seem rather attractive. In the case of the adrenaline–epinephrine argument, it appears that it was the Americans' turn to use their term. It will be interesting to see if all pharmacologists will be persuaded to come to terms with the fact that epinephrine binds to adrenoceptors (or perhaps they will have to go to speech therapists to learn how to pronounce 'epinephroceptors').

Such arbitrary prescriptiveness brings attempts to have useful and unambiguous nomenclature systems into disrepute, and, not surprisingly, upsets a number of scientists. Some current arguments concern whether thiamin or thiamine is the correct usage and whether or not NAD should be used instead of NAD^+. Although, it could be argued that there might be more useful activities for nomenclature and classification committees, there are a surprising number of active scientists who care deeply about such issues. In practice, the community is the ultimate arbiter, NAD (nicotinamide-adenine dinucleotide) displaced DPN (diphosphopyridine nucleotide), after a period of vituperative criticism from some people, because it was generally found to be superior. In contrast, the term heme-thiolate does not appear to have gained widespread popularity as an alternative to cytochrome *P*-450 and attempts to define enzyme activity in terms of the katal (mol product formed/s) rather than the unit (μmol product formed/min) have been largely unheeded. A lesson for nomenclature committees of all types must be that unless they perform functions that are generally perceived as being useful, they will be ignored. Nomenclature must be a service to the community not an end in itself.

A way of addressing differences between the nomenclature systems used by different groups, such as chemists, pharmacologists and biochemists, is the provision of a detailed glossary of alternative names of compounds and enzymes. In theory this should mean that, regardless of the name that you know a compound by, you should be able to find what you are looking for in the enzyme list. In practice, this has not been the case, due to space limitations associated with the printed version of the enzyme list. Because of the desire to keep the book form of the enzyme list to a manageable size, many of the alternative names for enzymes were removed from the list prior to publication of the 1992 edition. Many names that were ambiguous or misleading were also expunged. Although this was done with the best of intentions, it does not help people seeking information about an enzyme that they know by some name other than one of the approved names. Space limitations are not such a problem using the World Wide Web and the glossary is now being expanded as a look-up table to make finding information that much easier.

Rules and regulations

A detailed account of the rules for classification and nomenclature, the latest version of which was published in see Enzyme Nomenclature (1992, see http://www.chem.qmw.ac.uk/iubmb/ for more recent modifications and recommendations). These should make it a relatively simple matter to classify an enzyme correctly. However, there are some enzymes that do not fit into the general scheme. One such group is the peptidases (EC 3.4.–.–), since all of these can be regarded as catalysing the same reaction, hydrolytic cleavage of a peptide bond. Furthermore, it is often difficult to define their specificity as it depends on the nature of several amino acids around the peptide bond to be hydrolysed and the conformation of the substrate polypeptide chain. Therefore, their catalytic mechanism is also taken into account when classifying them. The peptidases are divided into two main sub-subclasses, the exopeptidases (3.4.11–3.4.19) and endopeptidases (3.4.21–3.4.24 and 3.4.99). The exopeptidases only act near the ends of polypeptide chains: if they act at the N-terminus to liberate a single amino acid, they are termed aminopeptidases (those that liberate two and three amino acids are called dipeptidyl-peptidases and tripeptidyl-peptidases, respectively). Those acting on peptide bonds that are one and two amino acids away from the C-terminus are carboxypeptidases and peptidyl-dipeptidases, respectively. The carboxypeptidases are further subdivided based on catalytic mechanism (serine-type carboxypeptidases, metallocarboxypeptidases and cysteine-type carboxypeptidases). Classification of the endopeptidases is based on the catalytic mechanism. Examples of the sub-subclasses are the serine endopeptidases, aspartic endopeptidases and metalloendopeptidases. Specificity, and source, are only used for naming the individual endopeptidases.

Having broken the rules once, it is of course easier to do so again, and there are several other cases where simple rules of nomenclature are inappropriate. The plethora of the restriction endonucleases occupies only three different spaces in the classification scheme:

EC 3.1.21.3: Type I site-specific deoxyribonuclease (also known as Type I restriction enzyme), which catalyses the endonucleophilic cleavage of DNA to give random double-stranded fragments with terminal 5′-phosphates; ATP is simultaneously hydrolysed.

EC 3.1.21.4: Type II site-specific deoxyribonuclease (also known as Type II restriction enzyme), which catalyses endonucleolytic cleavage of DNA to give specific double-stranded fragments with terminal 5′-phosphates.

EC 3.1.21.5: Type III site-specific deoxyribonuclease (also known as Type II restriction enzyme), which also catalyses endonucleolytic cleavage of DNA to give specific double-stranded fragments with terminal 5′-phosphates.

This is clearly not a great deal of help to those wishing to locate a specific restriction enzyme. However, the 'Comments' under each entry indicates that they comprise large groups of enzymes and, as well as providing limited further information on them refers to the REBASE (http://www.neb.com/rebase) restriction enzyme database that contains a complete, and regularly updated, list of these enzymes.

Unfinished business and work in progress

Several groups of enzymes need expansion, and advice from experts in these areas is needed. Such problems occur, for example, with the hydroxylases involving cytochrome *P*-450, which are classified as monooxygenases (EC 1.14.–.–), but where the nature of the 'physiological' substrate is often unknown, as well as with the DNA repair enzymes, and the protein kinases and phosphatases where the explosion of enzymes has made it difficult to keep abreast.

The treatment of multienzyme complexes and single polypeptides with several different activities merits further consideration. At present, these are listed under their separate activities, with comments that they exist as multienzyme complexes or multifunctional proteins. Thus, the multifunctional protein catalysing the synthesis of tryptophan is listed as its component activities (EC 2.4.2.18, 4.1.1.48, 4.2.1.20, 4.1.3.27 and 5.3.1.24), with the notes indicating that, in some organisms, the enzyme is part of a multifunctional protein, together with one or more of the other enzymes. This treatment has the advantages that it avoids uncertainties of where it should be classified and the fact that what is a multifunctional protein in some species may exist as separate components in another. Furthermore, it avoids the necessity for complex systematic names. For example, fatty-acid synthase (EC 2.3.1.85), which is an exception to this general rule, has the systematic name acyl-CoA:malonyl-CoA *O*-acyltransferase (decarboxylating, oxoacyl- and enoyl-reducing and thioester hydrolysing). An additional advantage of listing the components separately is that not all multifunctional proteins catalyse reactions that can be regarded as comprising a simple anabolic or catabolic reaction. Thus the combined reaction of the multifunctional 6-phosphofructo-2-kinase (EC 2.7.1.105) and fructose-2,6-bisphosphate 2-phosphatase (EC 3.1.3.46) would be simply that of an ATPase.

There are also decisions to be made as to how to include the ribozymes (small pieces of RNA with catalytic activity), catalytic antibodies (abzymes), novel activities arising from site-directed mutagenesis and the construction of chimeric proteins within the system, and whether or not it is appropriate to do so.

The way forward

In the 1958 edition of their book, Dixon and Webb stated that 'The number of enzymes is large—over 650—but not as large as perhaps might be expected in view of the complexity of the reactions occurring in living matter'. The number of enzymes classified by the Enzyme Commission has indeed risen steadily over the years, from 712 in 1964 to 3196 in 1992 (see Enzyme Nomenclature, 1965, 1973, 1979, 1992) for the steady development of the list of classified enzymes) and are is now approaching 3500 (see http://www.chem.qmw.ac.uk/iubmb/ for the most recent additions). More recent developments have involved changes in the information supplied as well as, of course, a steady expansion of the number of enzymes classified.

There are constraints on the NC-IUMB system; the nomenclature must accord with IUPAC rules and the information contained must be, as far as possible, correct and unambiguous. Furthermore, it comprises data that is based on function (reaction catalysed) and not structure. Thus, details of the kinetic behaviour of individual enzymes are not included and even information on cofactor requirements is often given in somewhat legalistic form. For example, a phrase like 'the enzyme from *E.coli* requires Mg^{2+}' leaves open the possibility that an enzyme catalysing the same reaction from another source may not require that metal ion.

Such considerations and the involvement of two committees, plus a number of advisory panels making recommendations to them on specific areas of nomenclature, means that the process of classifying and naming each new enzyme can be frustratingly slow. In the past, compilations of new enzymes were considered by the JCBN at an annual meeting, modified where appropriate and bought back to the next meeting for ratification before preparation as a supplement in the *European Journal of Biochemistry*. However, it has now proved possible to speed-up this process by the use of the internet for JCBN review and to pre-circulate material through the IUBMB home page (see http://www.chem.qmw.ac.uk/iubmb/) for comments from the community at large, before final ratification. It also means that the list contains a restricted amount of information. It is also possible to submit material proposing additional enzymes for inclusion or modifications for existing entries through that site or through the SWISS-PROT ENZYME site (see below).

Corteva Exhibit 2020 - Page 5 of 6
Inari v. Corteva - PGR2023-00022

However, the development of enzyme-related databases on the Web has transformed the situation and the Enzyme List forms the basis of several sites containing information on protein, gene or substrate structures, catalytic behaviour, metabolism, etc. The List can also be linked to sites providing more general information, databases containing such as thermodynamic databasesinformation, enzymes of relevance to clinical biochemists, more detailed references, etc. Two major enzyme databases that use information from the Enzyme List are SWISS-PROT ENZYME (http://www.expasy.ch/enzyme/) and the Kyoto University Ligand Chemical Database (http://www.genome.ad.jp/). Both of these also have useful links to other information of interest, such as metabolic pathway maps and BRENDA (an enzyme database that provides experimental data on each enzyme, as well as data that have been sanctioned by the Nomenclature Committee), as well as to the metabolite structural database KLOTHO (http://www.ibc.wustl.edu/klotho) and the Minnesota Biocatalysis/Biodegradation database (http://www.labmed.umn.edu/umbbd).

Perhaps if one were starting the enzyme classification system now, it might be considered worthwhile to adopt a system in which individual enzymes were assigned numbers that had no significance other than as a unique identifier for database manipulative purposes. The use of the information-rich EC number as a simple identifier makes some manipulative tasks somewhat more complicated. However, it is a fitting tribute to the insights and perseverance of Malcolm Dixon and Edwin Webb that the system that they devised and developed still stands us in such good stead over 40 years later.

Acknowledgements

We are grateful to the EU Biotechnology Programme (B104-CT-97-2099) and to NIH (grant no. GM56529-02) for funding of this work and to BioResearch Ireland for support. We also thank members of the JCBN for many helpful discussions.

References

Buchner,E. (1897) Alcoholic fermentation without yeast cells. *Ber. Dt. Chem. Ges.*, **30**, 117–124.

Cech,T.R. (1993) In Gesteland,R.F. and Atkins,J.F. (eds), *The RNA World*. Cold Spring Harbor Press, pp. 239–269.

Chapman,K.B. and Szostak,J.W. (1994) *Curr. Opin. Struct. Biol.*, **4**, 618–622.

Cornish-Bowden,A.(ed.) (1997) *New Beer in an Old Bottle: Eduard Buchner and the Growth of Biochemical Knowledge*. Universitat de València.

Dixon,M. and Webb,E.C. (1958) *Enzymes*. Longmans Green, London, pp. 183–227.

Dixon,M. and Webb,E.C. (1964) *Enzymes*. 2nd edn, Academic Press, New York.

Enzyme Nomenclature (1965) *Recommendations (1964) of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes, Together with their Units and the Symbols of Enzyme Kinetics (1965)*. Elsevier, Amsterdam.

Enzyme Nomenclature (1973) *Recommendations (1972) of the Commission on Biochemical Nomenclature on the Nomenclature and Classification of Enzymes, Together with their Units and the Symbols of Enzyme Kinetics (1973)*. Elsevier, Amsterdam.

Enzyme Nomenclature (1979) *Recommendations (1978) of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes (1979)*. Academic Press, New York.

Enzyme Nomenclature (1984) *Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes*. Academic Press, New York.

Enzyme Nomenclature (1992) *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, San Diego, CA.

Gopalan,V., Talbot,S.J. and Altman,S. (1994) In Nagai,K. and Mattaj,I.W. (eds), *RNA–Protein Interactions*. IRL Press, pp. 103–126.

Hofmann-Ostenhof,O. (1953) *Adv. Enzymol.*, **14**, 219.

Kühne,W. (1878) *Unters. A. d. physiol. Institut der Univ. Heidelberg*, **1**, 291.

Panico,R., Richer,J.-C. and Powell,W.H. (1994) *A Guide to IUPAC Nomenclature of Organic Compounds*. Blackwell Science, Oxford.

Payne,A. and Persoz,J.F. (1833) *Ann. Chim. (Phys.)*, **53**, 73.

BIOINFORMATICS

Vol. 16 no. 1 2000
Pages 41–47

The early introduction of dynamic programming into computational biology*

David Sankoff

Centre de recherches mathematiques, Universite de Montréal, CP 6128 succursale Centre-Ville, Montreal, Quebec H3C 3J7, Canada, E-mail: sankoff@ere.umontreal.ca

Introduction

In 1994–1995, DIMACS sponsored a theme year on computational biology. Among the numerous seminars and workshops was one organized by Alberto Apostolico and Raffaele Giancarlo, recapitulated in their 1998 paper, on the history and motivations for sequence comparison. In my participation in this event, I was led to consider some of the early interactions, at the Centre de recherches mathématiques (CRM) and elsewhere, in the field now known as computational biology. A short time earlier, I had read a 1989 paper by Walter Goad on the impact of Stanislaw Ulam in this field. The penultimate sentence in this article was a quote from Ulam himself 'I started all this', which puzzled me greatly. As far as I knew, Ulam had no impact in the early development of the field, and while he gave talks on it at least from 1971 (cf. Ulam, 1972), the distance he defined was already published (Levenshtein, 1965) and the problem he proposed had already been solved, for all intents and purposes, in the molecular biology literature (Needleman and Wunsch, 1970) and elsewhere (Vintsyuk, 1968). I had also read a joint interview of Ulam and Mark Kac by Feigenbaum (1982), and this led me to reflect on this misperception on the part of Ulam (and of Goad), and to crystallize the realization that ironically, Kac, his colleague of many years, had played a crucial, if indirect, role in the earliest development of the field, especially that associated with the Centre de recherches mathématiques (CRM). Despite the fact that Kac had no personal research interest in the field, his encouragement of a number of junior mathematicians, and his role, intentional or not, in bringing researchers together, recur as important influences in several aspects of computational biology, and explain why I dedicate this article to his memory.

In presenting some of these thoughts to the DIMACS workshop, focusing on the early 1970s, my understanding of this period was broadened by comments from a number of participants, particularly Jerrold Griggs and Pavel Pevzner, and clarified by the presentation immediately following, by Michael Waterman, who concentrated on the era starting in the late 1970s.

In this article, I will draw on my recollections of the earliest phases of the field to describe how certain fundamental ideas found their ways into the vernacular of the computational biologist.

Dynamic programming for sequence comparison

In the late summer of 1971, the sequence comparison problem was brought to my attention by my friend Robert Cedergren, a biochemist interested in the evolution of ribonucleic acid sequences. In discussions with him, we formulated this as the longest common subsequence problem, though at the time we called it maximum matching. I realized that the application of mathematical induction led to a simple recurrence relation for solving the problem. We consider two sequences of length m and n of terms from any alphabet. If the two sequences are $a(1), \ldots, a(m)$ and $b(1), \ldots, b(n)$, then using the notation a_i for the prefix sequence $a(1), \ldots, a(i)$ and $M(i, j)$ for the longest common subsequence of a_i and b_j,

$$M(i, j) = \text{Max} \begin{cases} M(i-1, j), \\ M(i, j-1), \\ M(i-1, j-1) + \Delta \end{cases}$$

where

$$\Delta = 1 \quad \text{if } a(i) = b(j) \quad \text{and} \quad \Delta = 0 \quad \text{if } a(i) \neq b(j)$$

under the initial conditions $M(i, 0) = M(0, j) = 0$. Then $M(m, n)$ is the length of the longest common subsequence, and all longest common subsequences may be found by a simple traceback routine on the matrix M: starting with matrix cell (m, n), we choose the predecessor cell (i, j) to be any one of the cells $(m-1, n)$, $(m, n-1)$, $(m-1, n-1)$ that satisfies

$$M(m, n) = M(m-1, n),$$

$$M(m, n) = M(m, n-1),$$

$$M(m, n) = M(m-1, n-1) + \Delta$$

respectively. We continue by finding a predecessor of (i, j), then its predecessor, and so on, until we reach

*This paper is dedicated to the memory of Mark Kac.

© Oxford University Press 2000

Corteva Exhibit 2020 - Page 6 of 6
Inari v. Corteva - PGR2023-00022