DNA pol research, with new DNA pol discoveries bound to occur in the near future.

Acknowledgements

We thank H. Pospiech, R. Smith, F. Grosse and S. Hasan for critical reading of the manuscript and for help with the artwork. U.H. has been supported by the Swiss National Science Foundation and the Kanton of Zürich. H.P.N. is supported by the Deutsche Forschungsgemeinschaft and the Boehringer Ingelheim Fund. J.E.S. is supported by the Academy of Finland. H.P.N. and U.H. are participating in the Training Mobility and Research grant ERBMRXCT CT970125. We apologize to those authors whose work could only be cited indirectly in recent reviews owing to editorial limitations.

References

1 Hübscher, U. and Thömmes, P. (1992) DNA polymerase ϵ, in search for a function. *Trends Biochem. Sci.* 17, 55–58
2 Burgers, P.M.J. (1998) Eukaryotic DNA polymerases in DNA replication and DNA repair. *Chromosoma* 107, 218–227
3 Bridges, B.A. (1999) DNA polymerases for passing lesions. *Curr. Biol.* 9, R475–R477
4 Waga, S. and Stillman, B. (1998) The DNA replication fork in eukaryotic cells. *Annu. Rev. Biochem.* 67, 721–751
5 Steitz, T.A. (1999) DNA polymerases: structural diversity and common mechanisms. *J. Biol. Chem.* 274, 17395–17398
6 Foiani, M. *et al.* (1997). The DNA polymerase α-primase complex couples DNA replication, cell cycle progression and DNA damage response.*Trends Biochem. Sci.* 22, 424–427
7 Voitenleitner, C. *et al.* (1999) Cell cycle-dependent regulation of human DNA polymerase alpha-primase activity by phosphorylation. *Mol. Cell. Biol.* 19, 646–656
8 Desdouets, C. *et al.* (1998) Evidence for a Cdc6p-independent mitotic resetting event involving DNA polymerase α. *EMBO J.* 17, 4139–4146
9 Holmes, A. and Haber, J.E. (1999) Double-strand break repair in yeast requires both leading and lagging strand DNA polymerases. *Cell* 96, 415–424
10 Wilson S.H. and Singhal, R.K. (1998) Mammalian DNA repair and the cellular DNA polymerases. In *DNA Damage and Repair* (Nickoloff, J. A. and Hoekstra, M. F., eds), pp. 161–180, Humana Press
11 Klungland, A. and Lindahl, T. (1997) Second pathway for completion of human DNA base excision repair: reconstitution with purified proteins and requirement of DNase IV (FEN1). *EMBO J.* 16, 3341–3348
12 Stucki, M. *et al.* (1998) Mammalian base excision repair by DNA polymerases delta and epsilon. *Oncogene* 17, 835–843
13 Dianov, G.L. *et al.* (1999) Role of DNA polymerase β in the excision step of long patch mammalian base excision repair. *J. Biol. Chem.* 274, 13741–13743
14 Plug, A.W. *et al.* (1997) Evidence for a role for DNA polymerase β in mammalian meiosis. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1327–1331
15 Wilson, T.E. and Lieber, M.R. (1999) Efficient processing of DNA ends during yeast nonhomologous end joining. *J. Biol. Chem.* 274, 23599–23607
16 Jonsson, Z.O. and Hübscher, U. (1997) Proliferating cell nuclear antigen: more than a clamp for DNA polymerases. *BioEssays* 19, 967–975
17 Aboussekhra, A. *et al.* (1995) Mammalian DNA nucleotide excision repair reconstituted with purified protein components. *Cell* 80, 859–868
18 Umar, A. *et al.* (1996) Requirement for PCNA in DNA mismatch repair at a step preceding DNA resynthesis. *Cell* 87, 65–73
19 Longley, M.J. *et al.* (1997) DNA polymerase δ is required for human mismatch repair *in vitro*. *J. Biol. Chem.* 272, 10917–10921
20 Giot, L. *et al.* (1997) Involvement of the yeast DNA polymerase δ in DNA repair *in vivo*. *Genetics* 146, 1239–1251
21 Francesconi, S. *et al.* (1993) Fission yeast with DNA polymerase δ temperature-sensitive alleles exhibit cell division phenotype. *Nucleic Acids Res.* 21, 3821–3828
22 Sugino, A. (1995) The yeast DNA polymerases and their role at the replication fork. *Trends Biochem. Sci.* 20, 319–323
23 D'Urso, G. and Nurse, P. (1997) *Schizosaccharomyces pombe* cdc20^{+} encodes DNA polymerase ϵ and is required for chromosomal replication but not for the S phase checkpoint. *Proc. Natl. Acad. Sci. U. S. A.* 94, 12491–12496
24 Zlotkin, T. *et al.* (1996) DNA polymerase epsilon may be dispensible for SV40- but not for cellular DNA replication. *EMBO J.* 15, 2298–2305
25 Pospiech, H. *et al.* (1999) A neutralizing antibody against human DNA polymerase e inhibits cellular but not SV40 DNA replication. *Nucleic Acids Res.* 27, 3799–3804
26 Aparicio, O.M. *et al.* (1999) Differential assembly of Cdc45p and DNA polymerases at early and late origins of DNA replication. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9130–9135
27 Kesti,T. *et al.* (1999) DNA polymerase ϵ catalytic domains are dispensible for DNA replication, DNA repair and cell viability, *Mol. Cell* 3, 679–685
28 Dua, R. *et al.* (1999) Analysis of the essential functions of the C-terminal protein/protein interaction domain of *Saccharomyces cerevisiae* pol e and its unexpected activity to support growth in the absence of the DNA polymerase domain. *J. Biol. Chem.* 274, 22283–22288
29 Navas, T. *et al.* (1995) DNA polymerase ϵ links the DNA replication machinery to the S phase check point. *Cell* 80, 29–39
30 Wang, Z. *et al.* (1993) DNA repair synthesis during base excision repair *in vitro* is catalysed by DNA polymerase ϵ and is influenced by DNA polymerases α and δ in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 13, 1051–1058
31 Friedberg, E.C. and Gerlach, V.L. (1999) Novel DNA polymerases offer clues to the molecular basis of mutagenesis. *Cell* 98, 413–416
32 Gibbs, P.E. *et al.* (1998) A human homolog of the *Saccharomyces cerevisiae* REV3 gene which encodes the catalytic subunit of DNA polymerase ζ. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6876–6880
33 Johnson, R.E. *et al.* (1999) hRAD30 mutations in the variant form of Xeroderma Pigmentosum. *Science* 285, 263–265
34 Masutani, C. *et al.* (1999) The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase η. *Nature* 299, 700–704
35 Johnson, R.E. *et al.* (1999) Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, pol η. *Science* 283, 1001–1004
36 Woodgate, R. (1999) A plethora of lesion-replicating DNA polymerases. *Genes Dev.* 13, 2191–2195
37 Nelson, J.R. *et al.* (1996) Thymine-thymine dimer bypass by yeast DNA polymerase ζ. *Science* 272, 1646–1649
38 Nelson, J.R. *et al.* (1996) Deoxycytidyl transferase activity of yeast REV1 protein. *Nature* 382, 729–731
39 Masutani, C. *et al.* (1999) Xeroderma pigmentosum variant (XP-V) correcting protein from HeLa cells has a thymine dimer bypass DNA polymerase activity. *EMBO. J.* 18, 34491–3501
40 Mc Donald, J.P. *et al.* (1999) Novel human and mouse homologs of *Saccharomyces cerevisiae* DNA polymerase η. *Genomics* 60, 20–30
41 Sharief, F.S. *et al.* (1999) Cloning and chromosomal mapping of the human DNA polymerase ϑ (POLQ), the eighth human DNA polymerase. *Genomics* 59, 90–96
42 Wagner, J. *et al.* (1999) The *dinB* gene encodes a novel *Escherichia coli* DNA polymerase (DNA pol IV) involved in mutagenesis. *Mol. Cell* 40, 281–286
43 Tang, M. *et al.* (1999) UmuD′$_2$C is an error-prone DNA polymerase, *Escherichia coli* DNA pol V. *Proc. Natl. Acad. Sci. U. S. A.* 96, 8919–8924
44 Radman, M. (1999) Enzymes of evolutionary change. *Nature*, 401, 866–868

NPS@: Network Protein Sequence Analysis

A large number of sequences are being generated by the various genome sequencing projects. One of the major challenges in the biocomputing field is to derive valuable information from these protein sequences. The first prerequisite in this process is to access up-to-date sequence and structure databanks (e.g. EMBL, GenBank, SWISS-PROT, Protein Data Bank; for a catalogue, see Ref. 1) maintained by several biocomputing centres, such as NCBI, EBI, EMBL, SIB and INFOBIOGEN. Ideally, sequences are analysed using a maximal number of methods on a minimal number of different Web sites. To achieve this, we developed a Web server called NPS@ (Network Protein Sequence Analysis, http://pbil.ibcp.fr/NPSA) that became available in 1998. NPS@ is the protein sequence analysis Web server of the *Pôle BioInformatique Lyonnais*, a group of biocomputing teams (http://pbil.univ-lyon1.fr), and provides the user with many of the most commonly used tools for protein sequence analysis.

The general flowchart describing relationships between biological sequence data and biocomputing methods available on the NPS@ web server is given in Fig. 1. The main feature of NPS@ is the interconnection of all methods gathered within a simple and user-friendly Web interface. Thus, it provides an easy way for protein sequence analysis and avoids the rather tedious cut-and-paste operations between sites. Therefore, on the same server the user can (1) search for homologous protein sequences; (2) constitute a subset of matching protein sequences; (3) perform multiple alignments; (4) make secondary structure predictions, then generate a consensus

 PII: S0968-0004(99)01540-6

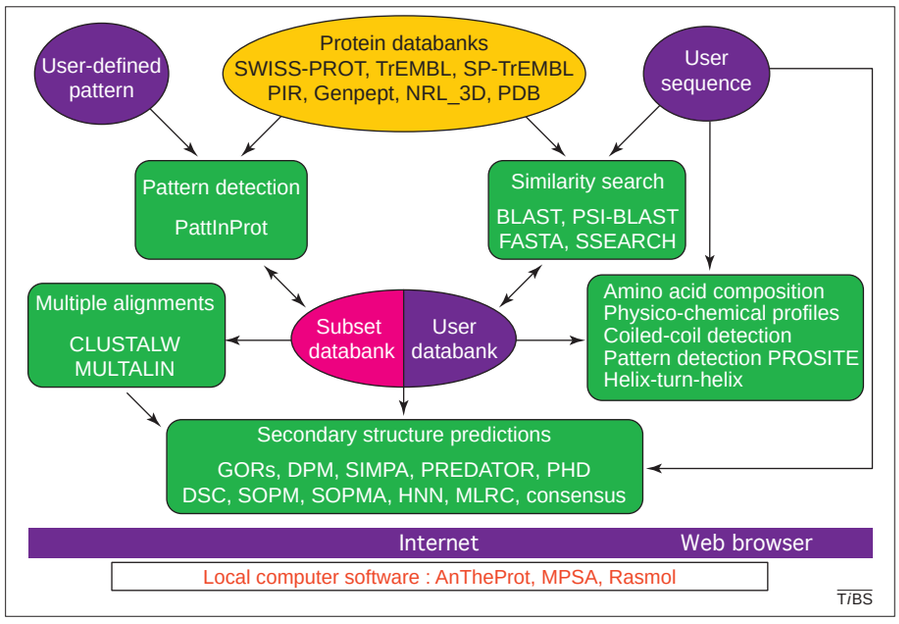

Figure 1

NPS@ flow chart. The user input of NPS@ can be either a user databank, a user-defined pattern or a single protein sequence (blue boxes). After performing pattern detection and similarity search on databanks (yellow box), a subset databank can be contructed (pink box). The subset databank can be edited and saved by the user as a new user databank that can be uploaded in NPS@. The user can apply various interconnected analysis tools (green boxes) to this databank.

structure and add them to the alignment; (5) plot physico-chemical profiles (hydrophobicity, antigenicity, potential membranous regions, predicted solvent accessibility); (6) detect functional sites or signatures (PROSITE definition[4]) specific for protein family; (7) predict the location of coiled-coil regions; (8) identify possible helix-turn-helix motifs. The user can access the server either with a single sequence, a personal databank or user-defined patterns (Fig. 1).

Single-sequence analysis

A typical example of a sequence similarity search is illustrated in Fig. 2. This search can be performed on several protein sequence databases (SWISS-PROT, SP-TrEMBL, TrEMBL, PIR, GenPept, NRL_3D) by using parallel versions of BLAST, PSI-BLAST (Ref. 2), SSEARCH and FASTA (Ref. 3). A common HTML interface (Fig. 2a) has been developed to display the similarity search results graphically. Such an interface is useful to analyse successive runs of PSI-BLAST quickly. In the results page, three links are provided for each subject sequence that allow the user to find biological information on the subject sequence by retrieving the database entry, checking the resulting alignment from the similarity search and analysing the subject sequence. Analysis is made possible by the special NPSA link (Fig. 2a) that opens a new HTML form (Fig. 2b), allowing access to the biocomputing tools available in NPS@. If the three-dimensional structure of the protein subject sequence is known (NRL_3D), several additional options allow the user to retrieve three-dimensional data from other biocomputing centres (SCOP, CATH, PDB) or to display the structure using viewer software such as RasMol (Fig. 2c).

The similarity search result page can also be used to construct a subset of related sequences for further analysis (Fig. 2a). In order to select the sequences, the expectation value (*E*), calculated using BLAST, is then used as a threshold for further selection. Moreover, the user can set the boundaries with respect to the query sequence to select subject sequences matching a particular region (e.g. a domain) of the query sequence. The user can use the graph at the top of the page to help choose these boundaries. All these choices are validated with the SELECT button. The user can also select or deselect some sequences with the checkboxes displayed at the beginning of each line. Once the selection has been made, the databank of selected sequences is generated with the EXTRACT button. Full sequences can be extracted from the query database. The user can also save partial sequences from the similarity-search-resulting alignment. In that case, the minimal number of residues in extracted sequences can be set. Finally, the query sequence can be added and the identical sequences can be removed from the user-made subset databank (Fig. 1).

Numerous analyses can be carried out on such a databank (Fig. 2d). First, the databank can be filtered either by a new similarity search (HOMOLOGY SEARCH button) or by several successive pattern scans (SEARCH PATTERN button) to detect protein sequences sharing the same sites or signatures. This pattern follows the well-known PROSITE syntax[4]. The APPLY NPS@ button allows the user to analyse each sequence with the tools seen in Fig. 2b. Another option is to align all protein sequences of the subset databank. This multiple alignment can be performed by using either a parallel version of CLUSTAL W (Ref. 5) or Multalin[6]. Several options are offered in the HTML result page of the alignment (Fig. 2e). The user can set the alignment width as well as a similarity level above which the residues are displayed and colour-coded. A consensus sequence is deduced and shown below the alignment. Moreover, this multiple alignment is coupled to protein secondary structure predicted by any combination of methods available on the NPS@ server SOPM (Ref. 7), SOPMA (Ref. 8), Hierarchical Neural Network and Multivariate Linear Regression Combination[9], Double Prediction[10], DSC (Ref. 11), GOR I (Ref. 12), GOR II (Ref. 13), GOR IV (Ref. 14), PHD (Ref. 15), PREDATOR (Ref. 16) and SIMPA96 (Ref. 17). A graphical plot of the predicted conformational states is given, as well as profiles for each conformational state. An interesting possibility is to download data, such as multiple alignments, predicted secondary structures and similarity-search result files, in protein-sequence-analysis software such as ANTHEPROT (Fig. 2f, http://www.ibcp.fr/ANTHEPROT) or MPSA (Fig. 2g, http://www.ibcp.fr/mpsa). For this purpose, the user has to configure the correct MIME types (chemical/x-antheprot or chemical/x-mpsa) in the Web browser(http://pbil.ibcp.fr/help/npsa_localviewing.html). These facilities are very useful to edit the alignment interactively (e.g. insert gaps, get amino acid positions in the alignment and in the sequence) and optimize it with respect to secondary structures and sites or signatures. The final alignment can be saved in MSF and ClustalW formats or exported as RTF or PostScript files.

In addition, ANTHEPROT and MPSA software are capable of submitting new analysis on the NPS@ server through client/server facilities (i.e. sequence data can be sent to the remote server via Internet for further analysis and results are sent back to the client's local software).

Personal databank

The second input point available is a personal databank that can be uploaded in NPS@ server from a local computer. On this databank, the user can apply all the various analyses described above and shown in Fig. 2d.

User-defined pattern

The third input point is user-defined patterns that are used in the PattInProt program that we have developed. This program scans a protein sequence or a databank for the presence of sites and signatures. The search can be performed with one or several mismatches or after setting a similarity level in order to detect more degenerate sequences. A recursive procedure allows for the successive searches of different patterns. This feature is particularly useful to generate and filter protein sequence databanks (see above).

Concluding remarks

A strong point and unique feature of this server is the coupling of secondary structure predictions with multiple alignments, performed on a sequence subset extracted following a similarity search (BLAST, PSI-BLAST, SSEARCH, FASTA or PattInProt). Indeed, all methods are interconnected in such a way that the output of a given analysis can be used as the starting point for another analysis. Efforts have been made to describe analytical methods and how to use the server through online help (http://pbil.ibcp.fr/help/help_npsaindex.html), news and reference topics that are accessible from a toolbar located at the top of each HTML page. This server is frequently and widely used (more than 1000 analyses per day). In conclusion, the main goal of this server was to assemble on a single site most of the tools used for protein sequence analysis, thereby reducing tedious cut-and-paste operations from many different sites. An automatic molecular modelling function of protein will be added in the near future. The NPS@ server provides an open platform that can be developed to accommodate future methodological advances and programs.

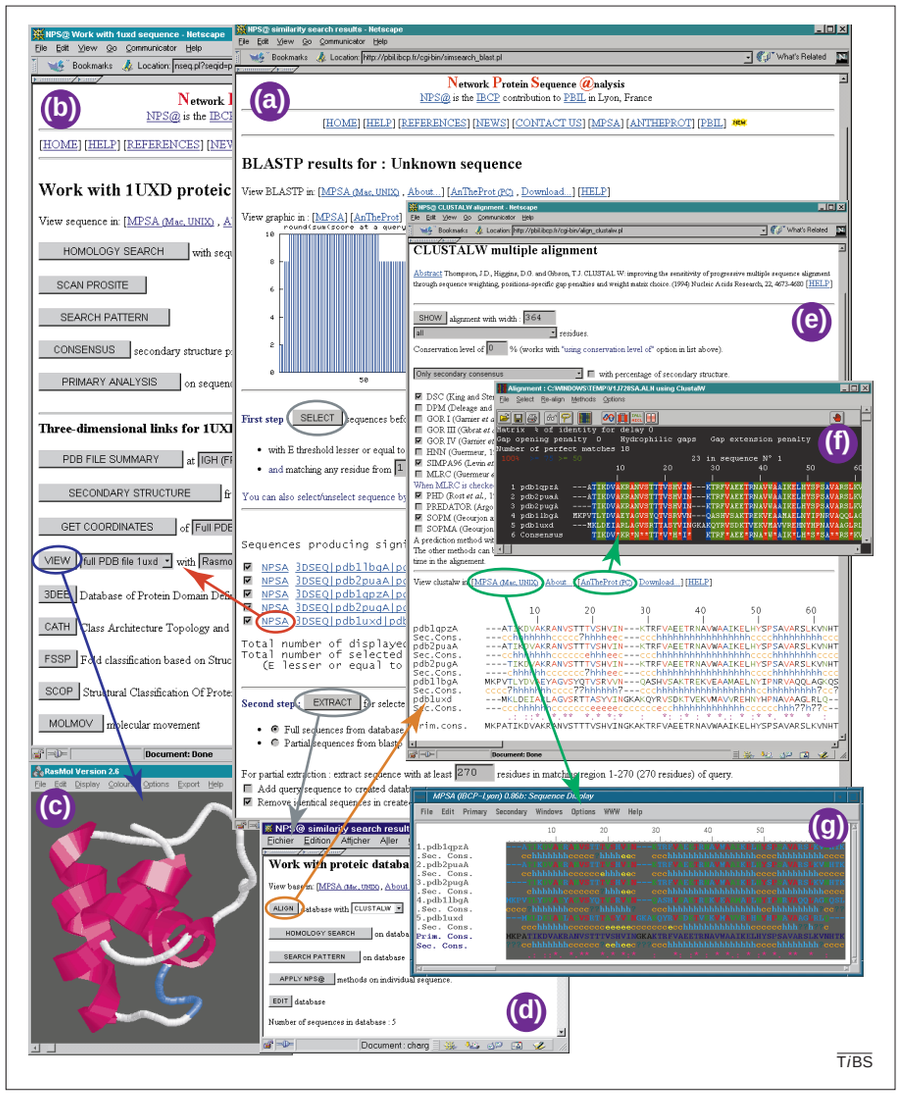

Figure 2

Snapshots of different stages in protein sequence analysis carried out with the NPS@ web server. **(a)** NPS@ output for BLASTP results; **(b)** the NPS@ link (red arrow) HTML page with tools and links available for individual sequence analysis; **(c)** the VIEW button (blue arrow) of the NPS@ link HTML page is useful to look at a PDB file in Rasmol (chemical/x-pdb MIME type); **(d)** SELECT and EXTRACT buttons (grey) permit to construct a subset databank and to display the intermediate form to work with it; **(e)** multiple alignment (orange arrow) coupled to secondary structure; **(f)** multiple alignment downloaded in AnTheProt software; **(g)** multiple alignment and secondary structures downloaded in MPSA software (green arrows).

Acknowledgements

The authors like to acknowledge financial support from CNRS, MENESR and Région Rhône-Alpes and thank all computing teams that have developed biocomputing methods for protein sequence analysis. C. Combet is the recipient of an ANRS doctoral fellowship. Thanks are due to D. Mandelman for textual improvements.

References

1 Kreil, D.P. and Etzold, T. (1999) DATABANKS – a catalogue database of molecular biology databases. *Trends Biochem. Sci.* 24, 155–157
2 Altschul S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database saerch programs. *Nucleic Acids Res.* 25, 3389–3402
3 Pearson, W.R. and Lipman D.J. (1988) Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
4 Hofmann *et al.* (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27, 215–219
5 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
6 Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* 16, 10881–10890
7 Geourjon, C. and Deléage, G. (1994) SOPM: a self-optimised method for protein secondary structure prediction. *Protein Eng.* 7, 157–164
8 Geourjon, C. and Deléage, G. (1995) SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction from multiple alignments. *Comp. Appl. Biosc.* 11, 681–684
9 Guermeur, Y. *et al.* (1999) Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics* 15, 413–421
10 Deléage, G. and Roux, B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* 1, 289–294

11 King R.D. and Sternberg, M.J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* 17, 3389–3407
12 Garnier, J. *et al.* (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120, 97–120
13 Gibrat, J.F. *et al.* (1987) Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* 198, 425–443
14 Garnier, J. *et al.* (1996) GOR secondary structure prediction method version IV *Methods Enzymol.* 266, 540–553
15 Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, 584–599
16 Frishman D. and Argos P. (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* 9,133–142
17 Levin J. (1997) Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.* 7, 771–776

C. COMBET, C. BLANCHET, C. GEOURJON AND G. DELÉAGE

Institut de Biologie et Chimie des Protéines, IBCP-CNRS UPR 412, Laboratoire de conformation des protéines, 7 Passage du Vercors, 69 367 Lyon Cedex 07, France.
Email: g.deleage@ibcp.fr

Macromolecular interactions: tracing the roots

In the 1990s there has been an explosion of articles in the biochemical literature focusing on various aspects of macromolecular interactions. Such heightened recognition of the significance of this subject has led some to hail the area as 'the new biochemistry'. In fact, this realm of biological research is quite 'old', with experimental and theoretical roots dating well over 50 years ago and with conceptual bases reaching back to the early 19th century. Revisionist history is a current (and controversial) literary development in the world of academia, which, when applied to biochemistry, seems to have resulted in the (re)interpretation of many of the discoveries in present-day biochemistry and cell biology. I believe that current researchers should be aware of the seminal ideas and findings of the various scientists of the past and so here, I present a brief history of work on macromolecular interactions.

Philosophical aspects

In philosophical terms, the significance of macromolecular interactions in cell biology is linked, ironically, to the notion of vitalism. Various debates concerning the role of a 'vital force' in characterizing the living state raged from the late 18th century well into the 20th century (see Ref. 1). The most extreme – and most (in)famous – vitalist position followed the late 18th-century 'design' argument that life is unique and cannot be explained as something acting physically *within* matter; life, in this view, must be accepted as a result of an imponderable 'force' operating *through* the system from *without*. This nonphysical conception led to the permanent (and negative) stigma commonly attached to the term 'vitalism' today. However, there was another belief that the expression 'vital force' simply connotes that ordinary physicochemical forces somehow act on the parts of the living being to achieve the organismal whole, following the Aristotelian concept that 'the whole is greater than the sum of its parts' (see Ref. 1). Today, 'complexity theory' has, indeed, solidified such a view of life (see Ref. 2). During the 19th century, though, this notion was advanced by the 'protoplasmic theory of life', which attempted to reduce the mechanistic definition of the living state to the cellular level and generated a variety of holistic physical models based on putative associations of the particulate matter within living cells (see Ref. 3).

In the late 19th century, the discovery of isolable enzymes, along with the finding by the Büchners that fermentation can occur in extracts of yeast cells, shifted attention away from the issue of the organization of cellular elements. There was, however, still discussion of organization versus nonorganization of enzyme action. For example, the eminent 19th-century physiologist Claude Bernard[4] differentiated '.... two kinds of fermentation, ... the one produced by .... an organized or *structured* ferment, the other produced by nonorganized ferments' [original italics].

The 'bag of enzymes' view of cells, nevertheless, held sway for the first half of the 20th century, as the reductionists pressed the philosophy that, if we know all the parts, we can conjure the whole – a conviction that is seriously flawed, if not completely wrong (see Ref. 5). Notably, many prominent biologists (Chambers, Cori, Krebs, Loeb, Peters, Wilson, to name a few) expressed the clear need to consider organizational ideas to explain differences between the observed behavior of cells and that of cell-free systems. For example, Krebs[6], in the 1930s, wrote that, 'Since all essential metabolic phenomena are bound to the cell structure, the tissue *Brei*, used so often in the past – in which the structure is destroyed – is unsuitable for metabolic investigations.' At about the same time, Rudolph Peters[7], arguing from theoretical grounds of physical chemistry, elaborated how, 'The view that is presented here differs from most others in the stress which is laid upon architecture. Its keynote is the complete (or nearly complete) structural organization of the cell. I believe this to be organized not only in respect of its grosser parts such as the nucleus, but also in regard to the actual chemical molecules of which it consists.'

Experimental beginnings

The demonstration of complete metabolic processes such as glycolysis, cholesterol synthesis and fatty acid synthesis in homogenized cell-free preparations was probably responsible for the belief that cellular integrity is essentially irrelevant to potential structural associations involved in metabolism. In the case of macromolecule synthesis, however, it was apparent early on that protein–protein, protein–nucleic acid and nucleic acid–nucleic acid interactions abound. The highly processive mechanisms observed in these pathways, catalysed by large proteinaceous complexes, implied the existence of almost perfect 'channeling' (i.e. molecular compartmentalization) of metabolic intermediates (see below).

David Green, in the late 1940s, reported the isolation of an aggregated system containing all of the enzymes of the Krebs tricarboxylic acid (TCA) cycle – which he dubbed the 'cyclophorase complex' (reviewed in Ref. 8). Green introduced the general term 'multienzyme complex' to designate 'an organized mosaic of enzymes in which each of the large number of component enzymes is uniquely located to permit efficient implementation of consecutive reaction sequences'. [More recently, I proposed the term 'metabolon' for such units of catalytic action (see Ref. 9).] It was

 PII: S0968-0004(00)01550-4