Multiple Sequence Alignment as a Workbench for Molecular Systems Biology

Julie D. Thompson* and Olivier Poch

Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Molculaire et Cellulaire, (CNRS/INSERM/ULP), BP 10142, 67404 Illkirch Cedex, France

Abstract: Recent progress in experimental techniques such as high-throughput genome sequencing, proteomics, transcriptomics and interactomics have lead to a new demand for integrated computational analyses, capable of systematically organizing these heterogeneous, fragmentary data into a coherent whole. As a consequence, novel system-level bioinformatics solutions are now being developed with the goal of understanding and predicting the behaviour of complex systems, such as molecular pathways, cells, tissues, organs and even whole organisms. Multiple alignments of both nucleotide and protein sequences play a central role in many of these applications, which range from the identification of genes and their products, *via* the characterisation of their 3D structure and their molecular and cellular functions, to the prediction of the phenotypic consequences of mutations, reverse engineering and drug design. In a multiple sequence alignment, structural and functional data can be combined with evolutionary information to allow reliable data validation, consensus predictions and rational propagation of information from known to unknown sequences. Clearly, integration at this scale calls for high quality, automatic multiple alignments. Alignment techniques are now responding to the challenge, with current developments moving away from a single all-encompassing algorithm towards more co-operative, knowledge based systems. However, the success of these methods relies on the efficient integration of information from different databases and the close cooperation of the different data mining and investigation algorithms. A large community effort is now underway to develop standards for data exchange and organisation that will facilitate collaborations between the various resources, in order to support improved domain understanding and to provide better decision-making systems and services for the biologist.

Keywords: Multiple sequence alignment, multiple alignment quality, systems biology, data integration.

I. INTRODUCTION

The field of molecular biology has been transformed by the availability of numerous complete genome sequences, together with the new information resources that are being created from the raw data produced by different high throughput technologies in fields such as transcriptomics, proteomics, or interactomics. This new wealth of information has opened up the possibility of system-level studies aimed at the elucidation, design or modification of complex structures, such as regulatory pathways, cells, tissues or even complete organisms. Systems biology aims to explain such complex biological systems through the integration of experimental data and computational research in order to formulate mathematical models that describe the structure of the system and its response to individual perturbations [1-4]. The goal is not simply to produce a catalogue of the individual components or even interactions, but to understand how the system components fit together, the effect of each individual part on its neighbours, and how various parameters such as concentrations, interactions, and mechanics change over time. Although a complete systems-level understanding of any biological process is still some way off, initial studies have begun to provide insights into how cellular networks may be organised [5]. For example, new biological pathways have been elucidated from analysis of protein-protein interaction networks derived from experimental methods such as yeast two-hybrid approach and mass spectrometry, in conjunction with other types of data, including protein function, subcellular location and gene expression profiles [6]. In addition to the definition of the system structure, system behaviour and response to external stimuli, such as environmental conditions, chemical injection or drug absorption are also being studied. For example, protein–protein interaction maps for the budding yeast *Saccharomyces cerevisiae* were combined with a genomic-scale data set describing the phenotypic role of all nonessential yeast proteins to study the recovery of the yeast from exposure to DNA-damaging agents [7]. A systematic integration of technologies is also being applied in the pharmaceutical industry to identify molecular functions and pathways associated with a disease and to improve the drug discovery pipeline, e.g. [8,9].

Systems modelling of molecular networks clearly requires an interdisciplinary approach, with input from genetics and molecular biology, chemistry, computer science and mathematics amongst others. It is a highly iterative process involving cycles of data collection, quantitative modelling, hypothesis formulation and testing, and model refinement. Clearly, new bioinformatics approaches are needed in order to manage and extract the important information from the mass of experimental or predicted data available today. In this context, multiple alignments of molecular sequences provide an ideal environment for the reliable integration of information from a complete genome

*Address correspondence to this author at the Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Molculaire et Cellulaire, (CNRS/INSERM/ULP), BP 10142, 67404 Illkirch Cedex, France; Tel: (33) 3 88 65 32 00; Fax: (33) 3 88 65 32 01; E-mail: julie@igbmc.u-strasbg.fr

1574-8936/06 $50.00+.00 © 2006 Bentham Science Publishers Ltd.

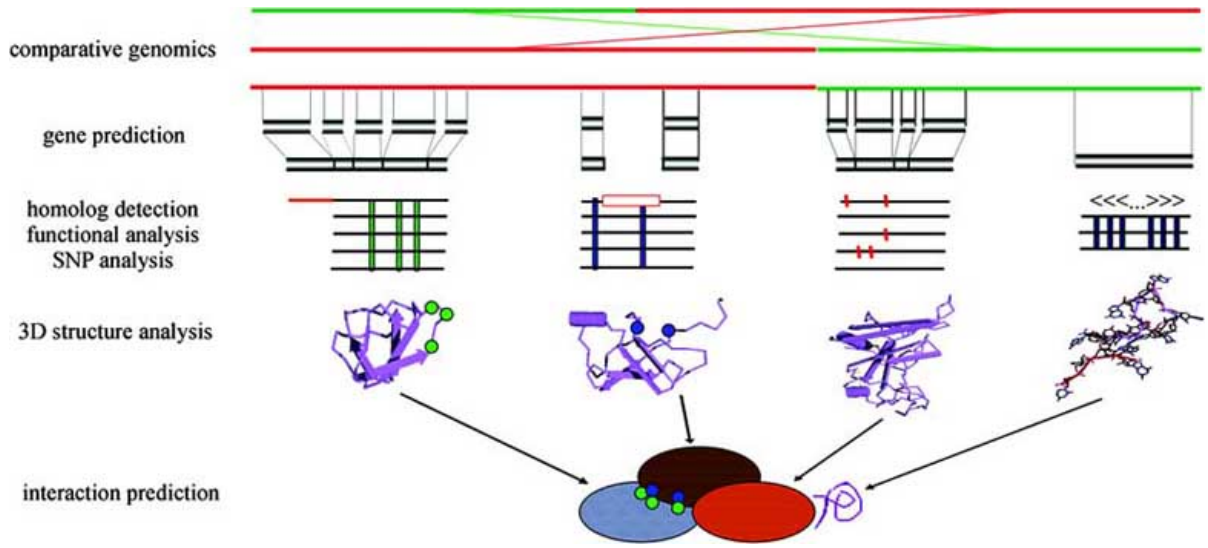


Fig. (1). Common multiple alignment applications: (i) comparison of complete genomes; syntenic regions are shown in red and green, (ii) gene structure prediction; exons are shown aligned on the genome for 3 protein coding genes and 1 RNA gene (iii) homolog detection by database searching, functional prediction and SNP analysis use information about conserved residues (shown as shaded bars on the multiple alignment), (iv) 3D structure analysis (conserved residues are indicated as coloured circles) (v) protein-protein and protein-RNA interactions.

to a gene and its related products. By placing the sequence in the framework of the overall family, multiple alignments can identify important structural or functional motifs that have been conserved through evolution, but can also highlight particular non-conserved features resulting from specific events or perturbations [10,11]. Multiple alignments thus allow reliable data validation, consensus predictions and rational propagation of information from known to unknown sequences, and provide a valuable workbench for integrated systems analysis, hypothesis generation and experiment-planning advice.

II. THE CENTRAL ROLE OF MULTIPLE ALIGNMENT IN THE POST-GENOMIC ERA

Multiple sequence alignment has become a fundamental tool in many different domains in modern molecular biology. One of the most well known applications is in evolutionary studies to define the phylogenetic relationships between organisms [12]. But multiple alignments are also exploited in numerous other tasks ranging from comparative multiple genome analysis to detailed structural analyses of gene products and the characterisation of the molecular and cellular functions of the protein. Fig. **1** shows a schematic representation of some of these applications, which are discussed in more detail below.

Comparative Genomics

Comparative genomics is a powerful discipline that is becoming more and more informative as genomic sequence data accumulate [13]. The major principles are straightforward. As genomes evolve, large-scale evolutionary processes, such as recombination, deletion or horizontal transfer, cause frequent genome rearrangements [14]. Comparative analyses of complete genomes present a comprehensive view of the level of conservation of gene order, or synteny, between different genomes, and thus provide a measure of organism relatedness at the genome scale, e.g. [15-18]. Examples of such analyses include comparisons among enteric bacteria [19] and between mouse and human [20]. In spite of these evolutionary rearrangements, the DNA sequences encoding the proteins and RNAs responsible for the functions shared between distantly related organisms, as well as the DNA sequences controlling the expression of such genes, should be preserved in their genome sequences. Conversely, sequences that encode proteins or RNAs responsible for differences between species will themselves be divergent. For example, a comparison of the genomes of yeast, worms, and flies revealed that these eukaryotes encode many of the same proteins, but different gene families are expanded in each genome [21]. A similar observation was made in a comparison of 16 complete archaeal genomes, where comparative genomics revealed a conserved core of 313 genes that are represented in all sequenced archaeal genomes, plus a variable 'shell' that is prone to lineage-specific gene loss and horizontal gene exchange [22].

Gene Identification and Characterisation

One of the major challenges in the post-genomic era is the complete, functional annotation of the genes coded by the sequenced genomes. Once a new genome sequence is made available, the first essential step in genome annotation is gene discovery and validation. The most widely used approach to detect protein coding regions consists of employing heterogeneous information from different methods, including codon usage bias and *ab initio* prediction of functional sites in the DNA sequence, such as splice sites, promoters, or start and stop codons. The reliability of these prediction methods depends critically on the quality of the underlying multiple alignments used to construct the consensus sequences or profiles that represent the various signals [23,24]. For prokaryotic genomes, over 95% of proteins can be successfully identified, e.g. [25], although the exact determination of the start site location remains

problematic because of the absence of relatively strong sequence patterns. Therefore, in the re-annotation of the *Mycoplasma pneumoniae* genome [26], multiple sequence alignments were used to propose N/C-terminal extensions to the original protein reading frame. The process of predicting genes in higher eukaryotic genomes is further complicated by several additional factors, including complex exon/intron organization, alternative splicing variants, and the sheer size of the genomic sequence [27]. It has been shown that comparison of the *ab initio* predicted exons with protein, EST, or cDNA databases can improve the sensitivity and specificity of the overall gene prediction, e.g. [28]. Unlike protein-coding genes, noncoding RNA gene sequences do not have strong statistical signals. Therefore, comparative sequence analyses are generally used to detect either known RNA motifs, e.g [29] or conservation patterns that may represent conserved secondary structures, e.g. [30]. But identifying genes alone will not be sufficient and there has been growing interest in alternative splicing as a mechanism for expanding the repertoire of gene functions. For example, a method has been developed to detect alternative splicing in expressed sequence data and to generate databases of alternative splicing relationships for the human, mouse and rat genomes [31]. Alternative splicing patterns are also represented by multiple alignments in databases such as the AsMamDB [32] or ASG [33].

The next step in the genome annotation process is the structural and functional analysis of the gene products. The most widely used approach is based on homology. The hypothesis is that homologous sequences i.e. sequences that have evolved from the same ancestor, often share the same 3D structure and have similar functions, active sites or binding domains. Thus, experimental information can be transferred from the known to the unknown protein, if homology is established. Homology-based methods generally begin with a search for similar proteins in the public sequence databases using tools such as BLAST [34] or FASTA [35]. The sensitivity of these database search methods has been significantly improved by exploiting information from multiple alignments e.g. PSI-BLAST [36], HMMSEARCH [37], SAM [38]. Homologous sequences can also be identified by searching protein family databases such as Interpro [39], where domains/motifs are generally represented by multiple alignments. Once sequence similarity to an experimentally-determined protein has been detected, accurate multiple alignments can provide a reliable basis for the transfer of information from the known protein to the unknown one, for example in both 2D and 3D structure modelling or in functional annotations [40-44]. As an example, Fig. **2** shows part of a multiple alignment of glutamine amidotransferase protein sequences detected by a BlastP search using the sequence q8dct2 as a query. The alignment highlights the catalytic triad of the class I GATase domain, as well as the residues differentially conserved in the two subfamilies of GMP synthases and anthranilate synthases. Reliable identification of functional motifs can be obtained by cross-validation of predictions for individual sequences in the context of the family. Thus, the presence of a DEAH-box subfamily ATP-dependent helicase signature motif (Prosite PS00690) in sequence TRPG_SULSO can be confirmed as false positive because the hit occurs in a non-conserved region of the alignment. Conversely, the class I GATase domain signature (PS00442) is supported by the existence of several similar predictions, and can thus be propagated to the other unpredicted sequences in the alignment. Finally, a potential frameshift in the open reading frame of the query sequence q8dct2 has been detected that may be a natural frameshift or may be the result of a sequencing error.

Even when no annotated homologues are available, the identification of sequences that are conserved in specific phyla can be useful information. For example, a set of novel virulence-associated genes that constitute novel targets for antimicrobial therapy were identified by comparing the genes of unknown function conserved among six human pathogens causing persistent or chronic infections in humans [45]. In addition, the analysis of the extent of sequence variation at a given position in a set of multiply aligned sequences can be used to identify functional sequence motifs [46-51] or to improve the reliability of other predictions, such as transmembrane helices [52], or subcellular localisation [53]. Patterns of conservation can also be used by *ab initio* methods to predict RNA structures [54], protein domain boundaries [55-57] and 2D [58] and 3D structures e.g. [59]. More detailed structural analyses also exploit the information in multiple alignments. For example, binding surfaces common to protein families have been defined on the basis of sequence conservation patterns and/or knowledge of the shared fold e.g. [60-64].

Interaction Networks

In the systems biology view of cellular function, each biological entity is seen in the context of a complex network of interactions. Powerful experimental techniques, such as the yeast two-hybrid system [65] or tandem-affinity purification and mass spectrometry [66], are now used to determine protein-protein interactions systematically. In parallel with these developments, a number of computational techniques have been designed for predicting protein interactions. The performance of the Rosetta method, which relies on the observation that some interacting proteins have homologues in another organism fused into a single protein chain, has been improved using multiple sequence alignment information and global measures of hydrophobic core formation [67]. A measure of the similarity between phylogenetic trees of protein families has also been used to predict pairs of interacting proteins [68,69]. Another approach involves quantifying the degree of co-variation between residues from pairs of interacting proteins (correlated mutations), known as the "*in silico* two-hybrid" method. For certain proteins that are known to interact, correlated mutations have been demonstrated to be able to select the correct structural arrangement of two proteins based on the accumulation of signals in the proximity of interacting surfaces [70]. This relationship between correlated residues and interacting surfaces has been extended to the prediction of interacting protein pairs based on the differential accumulation of correlated mutations between the interacting partners [71]. More recently, an alternative strategy has been proposed involving interaction networks derived from experiments in model organisms in order to obtain information about interactions that may occur between the orthologous proteins in different organisms [72].

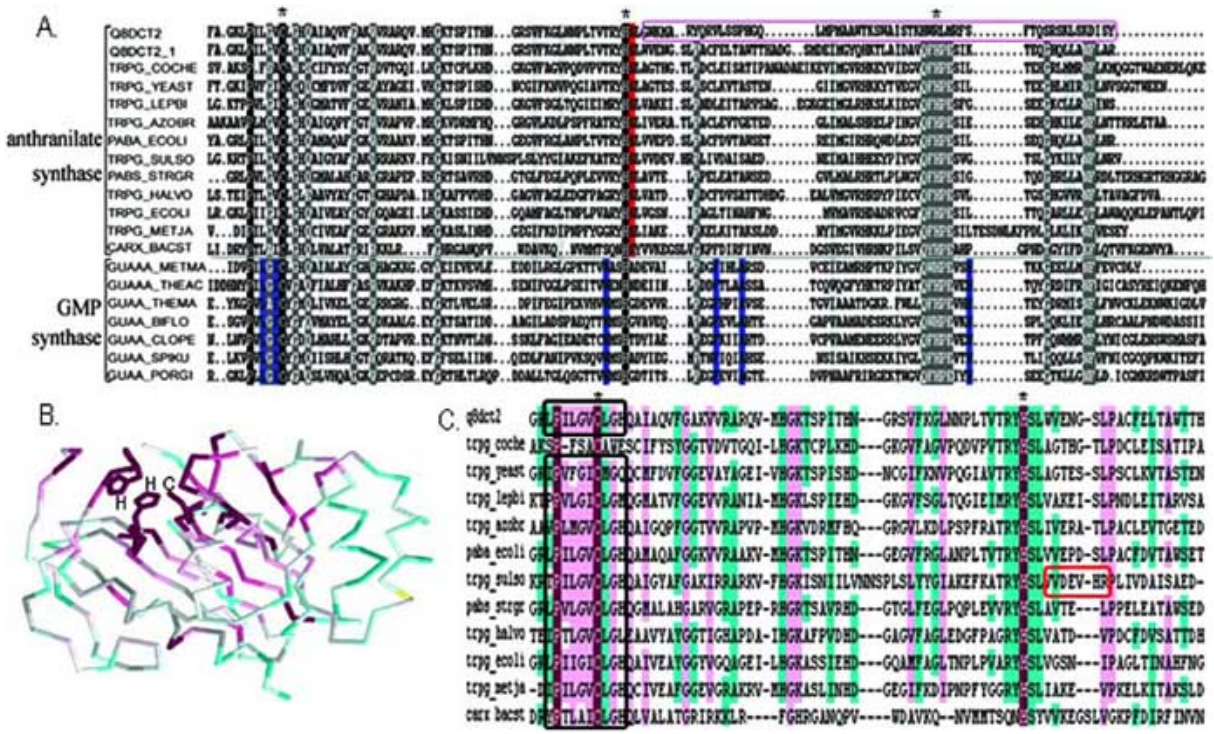


Fig. (2). A. Part of a multiple alignment of glutamine amidotransferase protein sequences. Active site residues are indicated by asterisks. Residues are coloured according to column conservation: black=100%; grey=80%; light grey=60%; red=100% in glutamine amidotransferase subfamily; blue=100% in anthranilate subfamily. The red box indicates a wrongly predicted sequence segment in sequence q8dct2. The corrected sequence is shown immediately below. B. The 3D structure of anthranilate synthase (PDB: 1I1Q) with residues coloured by CONSURF [141] (red=more conserved, blue=less conserved). Side chains are shown for the 3 labelled catalytic site residues. C. Detailed view of part of the alignment of the active site of anthranilate synthase. Residues are coloured according to their conservation (red=100%, pink=80%, blue=60%). Prosite motifs are shown as coloured boxes (black=Gatase_type_I, red=DEAH_ATP_helicase).

Genotype/Phenotype Correlations

A considerable effort is now underway to relate human phenotypes to variation at the DNA level. Most human genetic variation is represented by single nucleotide polymorphisms (SNPs) and many of them are believed to cause phenotypic differences between individuals [73]. One of the main goals of SNP research is therefore to understand the genetics of human phenotype variation and especially the genetic basis of complex diseases, thus providing a basis for assessing susceptibility to diseases and designing individual therapy. Whereas a large number of SNPs may be functionally neutral, others may have deleterious effects on the regulation or the functional activity of specific gene products. Non-synonymous single-nucleotide polymorphisms (nsSNPs) that lead to an amino acid change in the protein product are of particular interest because they account for nearly half of the known genetic variations related to human inherited disease [74]. With more and more data available, it has become imperative to predict the phenotype of a nsSNP *in silico*. Computational tools are therefore being developed, which use structural information or evolutionary information from multiple sequence alignments to predict a nsSNP's phenotypic effect and to identify disease-associated nsSNPs, e.g. [75,76].

III. EVOLUTION OF MULTIPLE ALIGNMENT ALGORITHMS

The multiple alignment problem has been widely studied for over twenty years and there now exists a very consequent literature describing the vast array of algorithms that have been exploited in the search for faster and more accurate alignments. A number of excellent reviews [77-79] have described in detail the introduction of multiple alignment based on an extension of the original pairwise dynamic programming algorithm [80] and the subsequent evolution of multiple alignment programs for both nucleic acid and protein sequences. Traditionally the most popular method has been the progressive alignment procedure [81]. A multiple sequence alignment is built up gradually by aligning the two closest sequences first and successively adding in the more distant ones. A number of alignment programs based on this method exist, notably ClustalW [82] and ClustalX [83] are based on the global Needleman-Wunsch algorithm [84]. In contrast, the Pima program [85] uses the Smith-Waterman algorithm [86] to find a local multiple alignment. Algorithms other than dynamic programming have also been exploited in the search for more accurate multiple alignments, including the use of Hidden Markov Models (HMM,s) in programs such as HMMER [37] or SAM [87], Genetic Algorithms in SAGA [88], segment-to-segment

alignments in Dialign [89], Gibbs Sampling [90] or iteration techniques, notably in the prrp program [91].

A number of studies have been performed recently to compare the various alignment methods and to evaluate the progress achieved in terms of both efficiency and alignment accuracy. A comparison of protein alignment programs [92], based on the benchmark alignment database, BAliBASE [93,94], showed that while global alignment methods in general performed better for sets of sequences that were of similar length, local algorithms were more successful at identifying the most conserved motifs in sequences containing large extensions and insertions. The same study also showed that the newer iterative methods often produced more accurate alignments, although these methods were generally slower than the progressive ones. Since then, a number of different 'gold standard' benchmarks have been proposed. For example, SABmark [95] provides sets of alignment problems covering the protein fold space, based on automatic pairwise structural alignments. An alternative method is used in PREFAB [96] where sequence sets are determined by PSI-BLAST [36] searches, although alignment accuracy is again assessed by comparison to an automatic pairwise structural alignment. The performance of some of these sequence alignment programs has also been evaluated using a benchmark of structural RNA multiple alignments [97]. It was shown that the 'twilight zone' of RNA alignment, below which sequence alignment methods are no longer reliable is in the 50–60% sequence-identity range. Below this limit, algorithms incorporating structural information, such as Dynalign [98], Foldalign [99], PMcomp [100] and Stemloc [101], outperformed pure sequence-based methods.

Although multiple alignment algorithms have improved significantly since their original introduction, a number of problems remain to be solved. Both the size and the complexity of the data sets that need to be routinely analysed are increasing. Large multi-domain proteins pose particular problems because of events such as domain duplications or domain recombinations that lead to non-linear alignments. Protein sequences that contain low-complexity regions, such as transmembrane segments also represent a difficult challenge for most of the existing alignment methods. It is clear that an accurate multiple alignment can no longer be constructed from the primary sequence data alone and there is general agreement that complementary algorithms and/or other information is necessary. Recent developments in multiple alignment methods have therefore tended towards integrated, knowledge-based systems incorporating additional information such as 3D structure, domain organisation or functional motifs. For example, the program DbClustal [102] uses information about locally conserved segments detected during a database search to guide or 'anchor' a global multiple alignment. T-Coffee [103] uses a library of pairwise alignments and constructs a multiple alignment with the highest possible level of consistency with the alignments within the library. MAFFT [104] and MUSCLE [96] both use fast algorithms to detect local sequence similarities, then use a progressive alignment algorithm together with an iterative refinement step to produce an accurate alignment for large sets of sequences. PROBCONS [105] is a Markov model-based progressive alignment algorithm that uses a probabilistic consistency measure to incorporate multiple sequence conservation information during pairwise alignment. Other methods incorporate 2D or 3D structure information to increase alignment sensitivity for both protein e.g. [106-110] and RNA sequences [111,112]. Another approach represents alignments by a graph structure and allows the alignment of protein sequences with shuffled and/or repeated domain structures [113,114]. In the near future, further advances in alignment quality will be likely with the inclusion of other experimental or predicted data, such as active sites, interactions, post-translational modifications, etc.

Alignment Quality Analysis

In the face of this complexity, the assessment of the quality and significance of a multiple alignment becomes a critical task. However, when a reference alignment is not available, an objective score is needed that describes the optimal or "biologically correct" multiple alignment. Several scoring systems have been proposed including the sum-of-pairs, minimum entropy and maximum likelihood scores. These measures, also known as objective functions, are currently used to evaluate and compare multiple alignments from different sources. They are also used in iterative alignment methods to improve the alignment by seeking to maximize the objective function. Some developments in this field have recently been reported, including global objective functions, e.g. [115-118] and measures of local reliability or column conservation e.g [119-121]. These methods are designed to identify the best alignment for a given set of sequences, however they do not provide any information about the extent of similarity existing between the different sequences. This particular problem has been addressed by a number of groups. For example, Errami *et al*. [122] analysed the agreement between predicted secondary structures of the aligned sequences to detect and discard unrelated sequences. Tress *et al*. [123] used sequence profiles generated from PSI-BLAST alignments to predict reliable regions between remotely related pairs of proteins. Thompson *et al*. [124] used profiles of locally conserved regions within subfamilies of the multiple alignment, and defined homologous regions based on an intermediate sequence analysis and the combination of individual weak matches to increase the significance of low-scoring regions.

IV. A RELIABLE WORKBENCH FOR SYSTEMS-LEVEL STUDIES

Thanks to the recent developments in multiple alignment algorithms, it is now possible to build accurate and reliable multiple alignments of large sequence sets, with the throughput time required by large scale projects. These multiple alignments provide an ideal environment for reliable integration of different kinds of information, such as alternative splicing, 3D structure, molecular interactions, cellular localisation, etc. As an example, Fig. **3** shows a multiple alignment of the vitamin D receptor (VDR) protein family, together with other related nuclear receptors. Vitamin D is critically important for the development, growth, and maintenance of a healthy skeleton from birth until death and mutations in the VDR gene are known to result in a generalized resistance to vitamin D, leading to the early onset of severe rickets [125]. The multiple alignment highlights the conserved N-terminal DNA binding domain

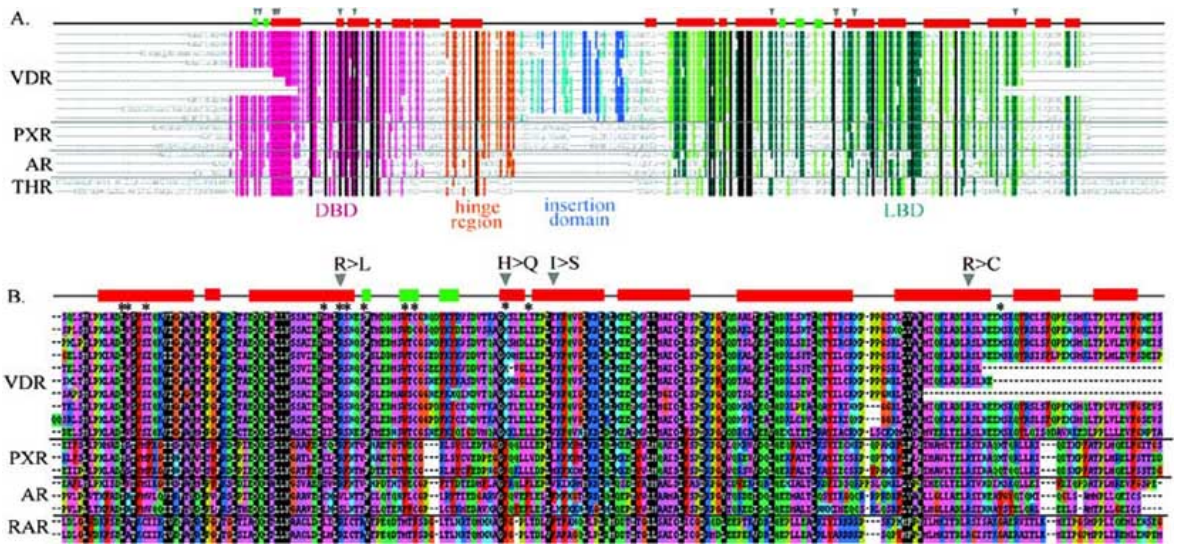


Fig. (3). A. Schematic representation of a multiple alignment of nuclear receptor protein sequences, divided into 4 sub-families. Different colours are used to shade the conserved columns in the 4 different domains (black = 100% conserved, dark shading = 60% conserved, light shading = 40% conserved). Secondary structure elements of human VDR (PDB: 1DB1) are shown above the alignment (red=helix, green=beta strand). B. Multiple alignment of the ligand binding domain (LBD). Alignment columns are coloured according to similar residue conservation (black=100% conserved residue or similar residue group (groups are DN;EQ;ST;KR;FYW;LIVM); orange=G, yellow=P, magenta=TSAC, cyan=NH, green=QED, red=FYW, pink=LIVM, blue=KR. Mutations for human VDR are indicated by black triangles above the alignment and are annotated as X>Y, where residue X is mutated to residue Y. Residue interactions with the ligand are indicated by asterisks above the alignment.

(DBD), a dispensable insertion domain (E1) and the ligand binding domain (LBD). The known mutations in the coding regions of the human VDR gene can be divided into two classes, representing two different phenotypes, based on this domain organisation. Mutations in the VDR DBD prevent the receptor from activating gene transcription, although vitamin D ($1,25D_3$) binding is normal. Patients with this DNA binding-defective phenotype do not respond to vitamin D treatment. In contrast, some patients with mutations in the LBD that cause reduced or complete hormone insensitivity have been partially responsive to high doses of calcium and vitamin D, although this often necessitates long term intravenous infusion therapy. For these patients, an alternative treatment using vitamin D analogs was proposed [126]. Knowledge of the 3D structure of the hormone-occupied VDR LBD [127] and the nature of the amino acid residues that contribute to the functional surface of the receptor allowed the selection of 3 candidate VDR mutations for analog treatment and the analogs that may be able to restore the functions of these mutants. The conserved residues arginine 274 and histidine 305 have been shown to contact the 1α-OH and 25-OH groups, respectively, in $1,25D_3$. The amino acid residue phenylalanine 251, which is localized to the E1 domain, is in a region that modulates Retinoid X receptor (RXR) heterodimerization and coactivator interaction. As a result, specific vitamin D analogs were identified that have the potential to interact with the receptor at amino acid contact points that differ from those utilized by the natural ligand, thus restoring the function of mutant VDRs. This example clearly illustrates the importance of polymorphism data that, combined with structural and evolutionary information, can form the basis for biochemical and cellular studies which may eventually lead to new drug therapies.

To achieve this level of integration, it is clear that data organisation and analysis techniques are now required to ensure that the pertinent information can be extracted and presented to the biologist in a clear, user-friendly format e.g. [128-132]. However, integration is currently hindered by syntactic and semantic differences between the different databases and applications, such as different naming conventions and terminology. The syntactic issue is now being addressed with the widespread adoption of standard file formats, such as the XML (eXtensible Markup Language) data exchange format [133]. To resolve semantic discrepencies, formal, structured vocabularies, known as ontologies are now being introduced in a number of areas for the management of biological knowledge [134]. In computer science, an ontology is defined as a formal, structured representation of the knowledge in a particular domain [135]. The most important aspect of an ontology is that it creates a shared understanding of a domain in a format that can be used by both humans and computers. Ontologies are thus used for automatic annotation of data, for the sharing of information from different resources and for the presentation of domain knowledge to researchers, and in particular to non-experts in the specific field. One of the most widely used bioinformatics ontologies is the Gene Ontology (GO) [136], which describes data about gene products. GO is composed of three separate hierarchical vocabularies, representing the function of a gene product, the process in which it plays a role and its cellular location. Numerous other ontologies are also being developed in other domains, including genome sequences [137], RNA sequences [138],

and molecular interactions [139]. In the field of multiple alignments, an ontology, MAO [140] has recently been introduced for both nucleic acid and protein sequences. One of the most powerful features of the MAO ontology is that it provides a natural, intuitive link between a number of different ontologies in the domains of genomics and proteomics, so that diverse experimental data and predicted information can be integrated in the context of the overall family alignment. Thus, structural and functional data can be combined with information about the conservation of the family and the variability observed at different residue sites.

V. PERSPECTIVES

The availability of fully sequenced genomes and the enormous amount of data from structural and functional proteomics projects has opened the way to new methods of analysing a gene's function, not only at the molecular level but also at the higher levels of the pathways, macro-molecular complexes, cells or organs the protein belongs to. In order to fully understand the functions and molecular interactions of a particular protein, such diverse information as cellular location, degradation and modification, 2D/3D structures, mutations and their associated pathologies, the evolutionary context and literature references must be assembled, classified and made available to the biologist. Integrated multiple alignments of complete sequences will provide an ideal workbench for the integration and presentation of the most vital and relevant aspects of all these sequence data.

Such an alignment network will facilitate information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and text searches, and sorting and visual exploration of search results. Although few, if any, of these methods have reached the status of validated proteomic tools, the rapid pace at which they are developing suggests that the rich and varied sources of information contained in the proteome will become increasingly accessible. The tools will allow the validation, visualization, integration and interpretation, in a biological context, of the vast amounts of diverse data generated by the application of proteomic and genomic discovery science tools. The potential applications are numerous, but will include such fields as the automatic annotation of the ever-increasing number of hypothetical proteins being produced by the high-throughput genome sequencing projects or the definition of characteristic motifs for specific protein folds. Hopefully, this will also have significant consequences for more wide-reaching areas, such as protein engineering, metabolic modelling, genetic studies of human disease susceptibility, and the development of new drug development strategies.

ACKNOWLEDGEMENTS

We would like to thank J.C. Thierry and D. Moras for their continued support. This work was supported by institute funds from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the Hôpital Universitaire de Strasbourg, and the Fond National de la Science (GENOPOLE).

REFERENCES

[1] Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Gene* **2001**; 2: 343-72.
[2] Kitano H. Computational systems biology. *Nature* **2002**; 420: 206-10.
[3] Kitano H. Systems biology: a brief overview. *Science* **2002**; 295: 1662-4.
[4] Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **2003**; 19: 551-60.
[5] Uetz P, Finley RL Jr. From protein networks to biological systems. *FEBS Lett* **2005**; 579(8): 1821-7.
[6] Cho S, Park SG, Lee do H, Park BC. Protein-protein interaction networks: from interactions to networks. *J Biochem Mol Biol* **2004**; 37: 45-52.
[7] Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* **2004**; 101: 18006-11.
[8] Davidov E, Holland J, Marple E, Naylor S. Advancing drug discovery through systems biology. *Drug Discov Today* **2003**; 8: 175-83.
[9] Apic G, Ignjatovic T, Boyer S, Russell RB. Illuminating drug discovery with biological pathways. *FEBS Lett* **2005**; 579: 1872-7.
[10] Woese CR, Pace NR. Probing RNA structure, function and history by comparative analysis, In: The RNA World, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY **1993**; 91-117.
[11] Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene* **2001**; 270: 17-30.
[12] Phillips A, Janies D, Wheeler W. Multiple sequence alignment in phylogenetic analysis. *Mol Phylogenet Evol* **2000**; **16**: 317-30.
[13] Hardison RC. Comparative genomics. *PLoS Biol* **2003**; 1: E58.
[14] Shapiro JA. A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering. *Gene* **2005**; 345: 91-100.
[15] Darling AE, Mau B, Blattner FR, Perna NT. GRIL: genome rearrangement and inversion locator. *Bioinformatics* **2004**; 20: 122-4.
[16] Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. *J Biomed Inform* **2002**; 35: 142-50.
[17] Elnitski L, Giardine B, Shah P, *et al*. Improvements to GALA and dbERGE II: databases featuring genomic sequence alignment, annotation and experimental results. *Nucleic Acids Res* **2005**; 33: D466-70.
[18] Ye L, Huang X. MAP2: multiple alignment of syntenic genomic sequences. *Nucleic Acids Res* **2005**; 33: 162-70.
[19] McClelland M, Florea L, Sanderson K, *et al*. Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res* **2000**; 28: 4974-86.
[20] International Mouse Genome Sequencing Consortium Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**; 420: 520–562.
[21] Rubin GM, Yandell MD, Wortman JR, *et al*. Comparative genomics of the eukaryotes. *Science* **2000**; 287: 2204-15.
[22] Makarova KS, Koonin EV. Comparative genomics of Archaea: how much have we learned in six years, and what's next? *Genome Biol* **2003**; 4: 115.
[23] Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* **2002**; 30, 4103-17.
[24] Qiu P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem Biophys Res Commun* **2003**; 309: 495-501.
[25] Aggarwal G, Ramaswamy R. Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J Biosci* **2002**; 27, 7-14.
[26] Dandekar T, Huynen M, Regula JT, *et al*. Re-annotating the mycoplasma pneumoniae genome sequence: adding value, function and reading frames. *Nucleic Acids Res* **2000**; 28: 3278-88.

[27] Zhang MQ. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet* **2002**; 3: 698-709.

[28] Bianchetti L, Thompson JD, Lecompte O, Plewniak F, Poch O. vALId: validation of protein sequence quality based on multiple alignment data. JBCB **2005**; (in press).

[29] Lowe T, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **1997**; 25: 955–64.

[30] Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* **2005**; 102: 2454-9.

[31] Grasso C, Modrek B, Xing Y, Lee C. Genome-wide detection of alternative splicing in expressed sequences using partial order multiple sequence alignment graphs. *Pac Symp Biocomput* **2004**;: 29-41.

[32] Ji H, Zhou Q, Wen F, Xia H, Lu X and Li Y. AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res* **2001;** 29: 260-3.

[33] Leipzig J, Pevzner P, Heber S. The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res* **2004**; 32: 3977-83.

[34] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol* **1990**; 215: 403-10.

[35] Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* **2000**; 132:185-219.

[36] Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**; 25: 3389-402.

[37] Eddy SR. Profile hidden Markov models. *Bioinformatics*. **1998**; 14: 755-63.

[38] Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **1999**; 14: 846-56.

[39] Mulder NJ, Apweiler R, Attwood TK, *et al*. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res* **2003**; 31: 315-8.

[40] Kunin V, Chan B, Sitbon E, Lithwick G, Pietrokovski S. (2001) Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs. *J Mol Biol* **2001**; 307: 939-49.

[41] Edwards YJ, Cottage A. Bioinformatics methods to predict protein structure and function. A practical approach. *Mol Biotechnol* **2003**; 23: 139-66.

[42] Michel F, Costa M, Massire C, Westhof E. Modeling RNA tertiary structure from patterns of sequence variation. *Methods Enzymol* **2000**; **317**: 491-510.

[43] Cozzetto D, Tramontano A. Relationship between multiple sequence alignments and quality of protein comparative models. *Proteins* **2005**; 58: 151-7.

[44] Frenkel-Morgenstern M, Voet H, Pietrokovski S. Enhanced statistics for local alignment of multiple alignments improves prediction of protein function and structure. *Bioinformatics* **2005**; 21: 2950-6.

[45] Garbom S, Forsberg A, Wolf-Watz H, Kihlberg BM. Identification of novel virulence-associated genes *via* genome analysis of hypothetical genes. *Infect Immun* **2004**; 72: 1333-40.

[46] Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **2001**; 17: 700-12.

[47] May AC. (2002) Definition of the tempo of sequence diversity across an alignment and automatic identification of sequence motifs: Application to protein homologous families and superfamilies. *Protein Sci* **2002**; 11: 2825-35.

[48] Rodi DJ, Mandava S, Makowski L. DIVAA: analysis of amino acid diversity in multiple aligned protein sequences. *Bioinformatics* **2004**; 20: 3481-9.

[49] Hoberman R, Klein-Seetharaman J, Rosenfeld R. Inferring property selection pressure from positional residue conservation. *Appl Bioinformatics* **2004**; 3: 167-79.

[50] Chen SW, Pellequer JL. Identification of functionally important residues in proteins using comparative models. *Curr Med Chem* **2004**; 11: 595-605.

[51] Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res* **2004**; 32: W424-8.

[52] Chen CP, Kernytsky A, Rost B. Transmembrane helix predictions revisited. *Protein Sci* **2002**; 11, 2774-91.

[53] Nair R, Rost B. Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins* **2003**; 53: 917-30.

[54] Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. *B M C Bioinformatics* **2004**; 5:140.

[55] George RA, Heringa J. SnapDRAGON: a method to delineate protein structural domains from sequence data. *J Mol Biol* **2002**; 316: 839-51.

[56] Rigden DJ. Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. *Protein Eng* **2002**; 15: 65-77.

[57] Nagarajan N, Yona G. Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics* **2004**; 20: 1335-60.

[58] Heringa J. Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr Protein Pept Sci* **2000**; 1:273-301.

[59] Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins* **1999**; Suppl 3: 177-85.

[60] Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **1996**; 257: 342-58.

[61] Armon A, Graur D, Ben-Tal N. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Bio* **2001**; 307**:** 447–63.

[62] Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* **2001**; 307**:** 1487–502.

[63] Qi Y, Grishin NV. PCOAT: positional correlation analysis using multiple methods. *Bioinformatics* **2004**; 20: 3697-9.

[64] Noivirt O, Eisenstein M, Horovitz A. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng* **2005**; 18: 247-53

[65] Ito T, Ota K, Kubota H, *et al*. Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol Cell Proteomics* **2002;** 8: 561-6.

[66] Gavin AC, Bosche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**; 415: 141-7.

[67] Bonneau R, Strauss CE, Baker D. (2001) Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins* **2001**; 43: 1-11.

[68] Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. *J Mol Biol* **2000**; 299: 283-93.

[69] Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **2001**; 14: 609-14.

[70] Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol* **1997**; 271: 511-23.

[71] Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **2002**; 47: 219-27.

[72] Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, Schneider MV. Comparative interactomics. *FEBS Lett* **2005**; 579: 1828-33.

[73] Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **2002**; 30: 3894-900.

[74] Stenson PD, Ball EV, Mort M, *et al*. Human gene mutation database (HGMD): 2003 update. *Hum Mutat* **2003**; 21: 577–81.

[75] Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* **2001**; 307: 683–706.

[76] Bao L, Cui Y. Prediction of the phenotypic effects of nonsynonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* **2005**; 21: 2185-90.

[77] Gotoh O. Multiple sequence alignment: algorithms and applications. *Adv Biophys* **1999**; 36: 159-206.

[78] Notredame C. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* **2002**; 3: 131-44.

[79] Batzoglou S. The many faces of sequence alignment. *Brief Bioinform* **2005**; 6: 6-22.

[80] Sankoff D. Minimal mutation trees of sequences. *SIAM J Appl Math* **1975**; 78: 35-42.

[81] Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **1987**; 25: 351-60.

[82] Thompson JD, Higgins DG, Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and matrix choice. *Nucleic Acids Res* **1994**; 22: 4673-80.

[83] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **1997**; 22: 4673-80.

[84] Needleman SB, Wunsch, CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **1970**; 48: 443-53.

[85] Smith RF, Smith TF. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng* **1992**; 5, 35-41.

[86] Smith TF, Waterman MS. (1981) Identification of common molecular subsequences. *J Mol Biol* **1981**; 147: 195-7.

[87] Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **1998**; 10: 846-56.

[88] Notredame C, Higgins DG. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res* **1996**; 24, 1515-24.

[89] Morgenstein B, Dress A, Werner T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci USA* **1996**; 93: 12098-103.

[90] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**; 262: 208-14.

[91] Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol* **1996**; 264: 823-38.

[92] Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res* **1999;** 27: 2683-90.

[93] Bahr A, Thompson JD, Thierry JC, Poch O. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res* **2001**; 29: 323-6.

[94] Thompson JD, Koehl P, Ripp R, Poch O. BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins* **2005**; 61: 127-136.

[95] Van Walle I, Lasters I, Wyns L. SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* **2005**; 7: 1267-8.

[96] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **2004**; 5: 113.

[97] Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* **2005**; 33: 2433-9.

[98] Mathews D, Turner, D. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences *J Mol Biol* **2002**; 317: 191–203.

[99] Hofacker I, Bernhart S, Stadler P. Alignment of RNA base pairing probability matrices *Bioinformatics* **2004**; 20: 2222–7.

[100] Hull HJ, Lyngsø R., Stormo G, Gorodkin J. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40% *Bioinformatics* **2005**; 21: 1815-24.

[101] Holmes, I. Accelerated probabilistic inference of RNA structure evolution *BMC Bioinformatics* **2005**; 6: 73.

[102] Thompson JD, Plewniak F, Thierry JC, Poch, O. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res* **2000**; 28: 2919-26.

[103] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **2000**; 302: 205-17.

[104] Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **2005**; 33: 511-8.

[105] Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* **2005**; 15: 330-40.

[106] Heringa J. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput Chem* **1999**; 23: 341-64.

[107] Jennings AJ, Edge CM, Sternberg MJ. An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein Eng* **2001**; 14: 227-31.

[108] Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* **2001**; 310: 243-57.

[109] O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* **2004;** 340: 385-95.

[110] Casbon J, Saqi MA. S4: structure-based sequence alignments of SCOP superfamilies. *Nucleic Acids Res* **2005**; 33: D219-22.

[111] Yang Q, Blanchette M. StructMiner: A Tool for Alignment and Detection of Conserved Secondary Structure. *Genome Inform Ser Workshop Genome Inform* **2004**; 15: 102-11.

[112] Jossinet F, Westhof E. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **2005;** in press.

[113] Grasso C, Lee C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics* **2004**; 20: 1546-56.

[114] Raphael B, Zhi D, Tang H, Pevzner P. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Res* **2004**; 14: 2336-46.

[115] Notredame C, Holm L, Higgins DG. COFFEE: an objective function for multiple sequence alignments. *Bioinformatics* **1998**; 14: 407-22.

[116] Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **1999**; 15: 563-77.

[117] Gonnet GH, Korostensky C, Benner S. Evaluation measures of multiple sequence alignments. *J Comput Biol* **2000**; 7: 261-76.

[118] Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. Towards a reliable objective function for multiple sequence alignments. *J Mol Biol* **2001**; 314: 937-51.

[119] Cline M, Hughey R, Karplus K. Predicting reliable regions in protein sequence alignments. *Bioinformatics* **2002**; 18: 306-14.

[120] Schlosshauer M, Ohlsson M. A novel approach to local reliability of sequence alignments. *Bioinformatics* **2002**; 18: 847-54.

[121] Beiko RG, Chan CX, Ragan MA. A word-oriented approach to alignment validation. *Bioinformatics* **2005**; 21: 2230-9.

[122] Errami M, Geourjon C, Deleage G. Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. *Bioinformatics* **2003**; 19: 506–12.

[123] Tress ML, Jones D, Valencia A. (2003) Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol* **2003;** 330: 705–18.

[124] Thompson JD, Prigent V, Poch O. LEON: multiple aLignment Evaluation Of Neighbours. *Nucleic Acids Res* **2004**; 32:1298-307.

[125] Malloy PJ, Pike JW. Feldman D. The Vitamin D Receptor and the Syndrome of Hereditary 1,25-Dihydroxyvitamin D-Resistant Rickets. *Endocrine Rev* **1999**; 20: 156-88.

[126] Gardezi SA, Nguyen C, Malloy PJ, Posner GH, Feldman D, Peleg S. A rationale for treatment of hereditary vitamin D-resistant rickets with analogs of 1 alpha,25-dihydroxyvitamin D(3). *J Biol Chem* **2001**; 31: 29148-56.

[127] Rochel N, Wurtz JM, Mitschler A, Klaholz B, Moras D. The crystal structure of the nuclear receptor for vitamin D bound to its natural ligand. *Mol Cell* **2000**; 5: 173-9.

[128] Hoon S, Ratnapu K, Chia J, *et al*. Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res* **2003**; 13: 1904-15.

[129] Rowe A, Kalaitzopoulos D, Osmond M, Ghanem M, Guo Y: The discovery net system for high throughput bioinformatics. *Bioinformatics* **2003**; 19: 225-31.

[130] Johnson JM, Mason K, Moallemi C, Xi H, Somaroo S, Huang ES. Protein family annotation in a multiple alignment viewer. *Bioinformatics* **2003**; 19: 544-5.

[131] Shah SP, He DY, Sawkins JN, *et al*. Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinformatics* **2004**; 5: 40.

[132] O'Donoghue SI, Meyer JE, Schafferhans A, Fries K. The SRS 3D module: integrating structures, sequences and features. *Bioinformatics*. **2004**; 20: 2476-8.

[133] Achard F, Vaysseix G, Barillot E. XML, bioinformatics and data integration. *Bioinformatics* **2001**; 17: 115-25.

[134] Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet* **2004**; 5: 213-22.

[135] Gruber TR, Toward Principles for the design of ontologies used for knowledge sharing. In: Formal Ontology in Conceptual Analysis and Knowledge Representation. Kluwer Academic Publishers, Deventer, The Netherlands 1993; 199-220.

[136] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res* **2001**; 11: 1425-33.

[137] Eilbeck K, Lewis SE, Mungall CJ, *et al*. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* **2005**; 6: R44.

[138] Leontis N, Berman H, Brenner S, *et al*. The RNA Ontology Consortium: An open invitation to the RNA community. *RNA* **2005**; submitted.

[139] Hermjakob H, Montecchi-Palazzi L, Bader G,*et al*. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol* **2004**; 22, 177-83.

[140] Thompson JD, Holbrook SR, Katoh K, *et al*. MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. *Nucl Acids Res* **2005**; 33: 4164-4171.

[141] Landau M, Mayrose I, Rosenberg Y, *et al*. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* **2005**; 33:W299-302.

Received: July 13, 2005 Revised: August 29, 2005 Accepted: August 29, 2005