REVIEWS

Adv Clin Exp Med 2007, **16**, 1, 85–93
ISSN 1230-025X

MICHAŁ PIAST, IRENA KUSTRZEBA-WÓJCICKA, MAŁGORZATA MATUSIEWICZ,
MAŁGORZATA KRZYSTEK-KORPACKA, TERESA BANAŚ

Bioinformatics: From arduous beginnings to molecular databases

Bioinformatyka: od trudnych początków do molekularnych baz danych

Department of Medical Biochemistry, Silesian Piasts of Medicine in Wrocław, Poland

Abstract

This is a brief review of the origins of bioinformatics and the development of biological databases and molecular analysis tools. The paper covers the period from 1945 (Sanger's work on insulin) to 2004 (introduction of the latest MEGA version, GenBank release 143). For the purpose of this article, the term "bioinformatics" means the discipline involving biology and computer science, and the term "computational biology" is understood as a process of biological data analysis and interpretation (**Adv Clin Exp Med 2007, 16, 1, 85–93**).

Keywords: Software Databases, Evolution, Phylogenetics, Sequence alignment

Streszczenie

Zwięzły opis początków bioinformatyki, rozwoju molekularnych baz danych i narzędzi do analizy sekwencji aminokwasowych i nukleotydowych. Praca obejmuje okres od roku 1945 (prace Sangera nad insuliną) do roku 2004 (wprowadzenie najnowszej wersji programu MEGA, aktualizacja bazy GenBank do wersji 143). Na potrzeby niniejszej pracy określenie „bioinformatyka" oznacza dyscyplinę łączącą biologię i nauki komputerowe, a termin „biologia obliczeniowa" jest rozumiany jako proces analizy i interpretacji danych biologicznych (**Adv Clin Exp Med 2007, 16, 1, 85–93**).

Słowa kluczowe:

History of Bioinformatics

It is no easy task to trace back to the very beginning of the marriage of biology and computational machines. When asked about the origins of bioinformatics, most of us would answer that it started in the era of the Internet and genome sequencing projects. Yet computational machines were applied in molecular biology years before DNA sequencing began. In the early 1960s the amount of data from protein chemistry begun to grow rapidly. Scientists realized that to cope with this problem they needed to combine molecular biology, mathematics, and computers. It seemed almost impossible to analyze the growing collection of protein data without the help of powerful computational devices. These machines became available to scientists in the decade following the Second World War, but by the late 1960s not all biologists had access to them [1].

One of the golden periods in computational biology was when Frederick Sanger successfully sequenced insulin during the decade from 1945 to 1955 [2]. Sanger received the Nobel Prize in 1958 and established the polypeptide theory of protein structure. It took Sanger's group a whole decade to sequence a complete protein. The process would have probably lasted longer but for Sanger's great knowledge and experience in protein degradation. Although valuable, Sanger's method was also "mind-numbing"; biochemists thus developed other, less time-consuming techniques, e.g. Edman's degradation, ion exchange columns, and electrophoresis, all of which made sequencing more efficient. It was a matter of few years and amino-acid identification became automated.

Fig. 1. Margaret Dayhoff. Pioneer of bioinformatics. Author of *Atlas of Protein Sequences*. (Source: http://www.dayhoff.cc.)

Ryc. 1. Margaret Dayhoff – pionierka bioinformatyki, autor „Atlas of Protein Sequences" (Źródło: http://www.dayhoff.cc.)

Fig. 2. Russell Doolittle. Research Professor at the Center for Molecular Genetics, University of California, San Diego. (Photograph courtesy of Russell Doolittle)

Ryc. 2. Russell Doolittle – profesor Centrum Genetyki Molekularnej, Uniwersytet Kalifornijski, San Diego (Zdjęcie dzięki uprzejmości Russella Doolittle)

Moore and Stein made use of semi-automated methods and sequenced 124 amino acids of ribonuclease in half the time that Sanger had spent to work out the 51-amino-acid-long insulin molecule. Soon Edman put a fully automated machine into service performing the degradation reaction. The arrival of this machine set off an enormous growth of protein data. Widespread use of the automated techniques accelerated the development of amino-acid libraries. Computers were still almost of no use in biochemical laboratories, with the exception of John Kendrew's, who used a computer to predict the three-dimensional structure of myoglobin [1].

In 1957, IBM introduced the first high-level programming language, FORTRAN (FORmula TRANslation). This application was useful for scientific purposes and, what is more, it was relatively easy to learn. FORTRAN drove the development of computational biology and inspired the pioneering work of Margaret Dayhoff (Figure 1). During the 1960s she focused her attention on molecular evolution and became an associate director of the National Biomedical Research Foundation, whose aim was to speed up the development of computer software. Dayhoff succeeded in writing FORTRAN programs which were applied to determine amino-acid sequences in proteins. Using peptide fragments as an input data, she predicted all possible sequences that were reliable according to the data. Within a few minutes the program deduced the proper sequence of ribonuclease, an accomplishment that took Stein and Moore five years. When gene sequences became available, similar programs were developed by other biologists and used to analyze nucleid acids. Sequence analysis software had a crucial influence on developing libraries subsequently used in molecular evolution. Finally, Dayhoff founded the *Atlas of Protein Sequences*, a publication which endeavored to contain all known amino-acid sequences. The *Atlas* was the first known molecular biology database and a vital resource for computational scientists. In time it evolved into the Protein Information Resource (PIR). Again, computers proved to be a invaluable tool in the hands of scientists [1].

Computational machines were successfully used as an important instrument for studying phylogenetics. Because drawing all possible phylogenetic trees is virtually impossible if done manually, this was a field where computers, with their computational power, shone. A classic example is the work of Russell Doolittle (Figure 2), who began phylogenetic analysis by hand, subsequently turning his gaze to research programs. He made use of a DEC PDP11 mini-computer and home-grown programs to search sequences in an effort to find evolutionary and other biological relationships [3, 4].

Milestones in computational biology:

1953 – Insulin sequenced by Frederick Sanger

1957 – FORTRAN introduced

1965 – Atlas of Protein Sequences by Margaret Dayhoff

1980 – Review of computational biology in *Science*

1982 – GenBank freely accessible

1987 – Multiple sequence alignment
1988 – National Center for Biotechnology Information created
1995 – First bacterial genomes fully sequenced

Another case was the phylogenesis of cytochrome c, a protein whose amino-acid sequence was known from a variety of species. Fitch and Margoliash made use of programs similar to those applied by Dayhoff to make pair-wise comparisons among homologous amino-acid sites in 20 species. The program computed the minimum number of steps to change one cytochrome into another. After that, a phylogenetic tree was drawn, starting from a simple one with three branches, to more complicated ones involving other species. However, the program had some limitations, as it did not perform complete analysis and some work had to be done by hand. The computer drew several trees according to data set by the researcher, discarding the less accurate, so it was the scientist's decision to choose the most suitable one [1].

Early phylogenetic analysis was based on the statement that comparable proteins were homologous. Cytochrome c from various species definitely has a common ancestor, but in the case of distantly related proteins, detecting homology became a serious problem for pioneer molecular evolutionists. A solution came in the form of computer algorithms constructed especially to establish sequence homology. Fitch's calculations were based on alignments of selected sequences of two proteins. For each comparison the program estimated the minimum number of mutations needed to convert one sequence into the other. This method, studied further by others, especially Needleman and Wunsch, became one of the standard procedures in sequence alignment. The idea of alignment is so common because lining up sequences and comparing their residues is the easiest way to check sequence similarity. A comparable scheme adopted by Dayhoff used a mutation data matrix to predict the probability of substituting one amino acid with any other. The matrices soon became standard devices in molecular evolution analysis and influenced the development of more sophisticated tools [1, 5].

Computer programs and sequencing machines affected DNA experiments. As with proteins, the first genes were cloned and sequenced manually, but the whole process was drastically changed when sequenators were implemented. After that, gene libraries started to grow at such a rate that they are now doubling in size every nine months. Other specialized software allowed predicting gene-coding regions within larger genomes [5]. Implementing computers and specialized software in molecular biology and phylogenetics soon created a large amount of records. The first results of computational analysis were skeptically overlooked by "conservative" scientists, but with time they became a significant part of research. During the early years of bioinformatics, scientists had difficulties in finding appropriate journals because their work was often considered insignificant. The first bioinformatics papers began to appear in traditional biological journals such as *Gene* and the *Journal of Molecular Biology*. In 1980, *Science* published one of the first serious reviews of computational biology (Gingeras and Roberts). *Nucleic Acids Research* played a major role. *NAR* encouraged the publication of papers on DNA sequences, but it also acted as a forum for the newly developed union of computer science and computational biology. *NAR* now has a leading position amongst journals publishing papers devoted to sequence databases. The first issue of every year consists entirely of papers describing biological databases. Nowadays, authors can choose amongst many journals specializing in computational papers, e.g. *Bioinformatics*, *Journal of Computational Biology*, *Journal of Theoretical Biology*, and *Mathematical Biosciences* [6, 7].

Molecular Databases

Bioinformatics arose as a fusion of two trends in biology: the application of computer programs to the analysis of protein and nucleotide sequences and the storage of molecular sequences in computer databases. The history of databases started in 1965 with the work of Margaret Dayhoff and her *Atlas of Protein Sequences*, which became the foundation for the first online database, the Protein Information Resource. Sequence information from laboratories all around the world accumulates exponentially. To store and organize such tremendous amounts of data, sequence databases were developed. Most of these data are maintained by three major databases: GenBank, supported and distributed by the National Center for Biotechnology Information (NCBI), the Protein Identification Resource, maintained by the National Biomedical Research Foundation (NBRF), and the EMBL/SwissProt databases maintained by the European Molecular Biology Laboratory [8, 9].

A typical database contains several long ASCII text files with various information related to the particular sequence. To bind this information, scientists use binary files that provide indexing functions. Protein databases are usually divided into subdivisions based on the level of a sequence's annotation, i.e. the origin of the pro-

tein, author information, journal citations, etc. [8]. Nucleic acid databases are split into subcategories based on taxonomy and data type. The information gathered in molecular databases – sequences, annotations, taxonomy – is an invaluable tool in studies on molecular evolution [10–12]. Below, one will find short descriptions of a few of the largest sequence databases.

GenBank

GenBank is a public sequence database which stores DNA sequences of over 165,000 different species. An average of 1700 new species are being added to GenBank every month. Data in GenBank are permanently exchanged with the DNA Data Bank of Japan and the EMBL Data Library, thus ensuring worldwide coverage. It was built by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) of the US National Institutes of Health (NIH). Sequences come mainly from the authors' direct submission and bulk submission of ESTs (Expressed Sequence Tags). Another source of sequences is the US Office of Patents and Trademarks. A large part of the individual submissions (about 30%) is received through BankIt, a web-based submission tool. All sequence information, including annotations, are given directly by the authors. This makes sequence submission as simple as possible [13].

In August 2004 the database contained 41.8 billion nucleotide bases from over 37.3 million different sequences. Large parts of GenBank are complete genomes from over 180 bacterial and 20 eukaryotic organisms. GenBank grows at a very fast rate, doubling its data every 15 months (Figure 3). Such fast growth is possible because of the high capacity of ESTs (over 69% of all sequences). ESTs are a major source of new sequences and genes, their number increasing rapidly especially in five organisms: human (5.7 million records), rat (617,000 records), mouse (4.2 million records), *Ciona interstinalis* (684,000 records), and *Danio reiro* (575,000 records). Each sequence is described in detail. Basic information like scientific name, taxonomy, and bibliographic reference is followed by a more precise description, i.e. the coding and repeat regions, mutations sites, etc. GenBank sequences are divided into "divisions" corresponding to taxonomy, e.g. plants or primates. Large divisions are divided into smaller ones, as this makes file transfer more efficient. Every sequence has its own unique accession number, which is not subject to change even when there is a change in the sequence or annotation. NCBI assigns a "gi" number to each sequence to prevent mistakes when identifying specific sequences from different sources. When a sequence is changed, the new version is assigned its own "gi" number. This unique number is also given to protein-sequence translations.

All data are freely available over the Internet (http://www.ncbi.nlm.nih.gov). A common way of retrieving information from GenBank is via Entrez, an integrated database retrieval system. Not only sequences, but also NCBI taxonomy, Molecular Modeling Database (MMNDB), and MEDLINE references can be obtained through Entrez. GenBank is also released in compressed flat-file format accessible through anonymous file transfer protocol (ftp) (ncbi.nlm.nih.gov) [13].

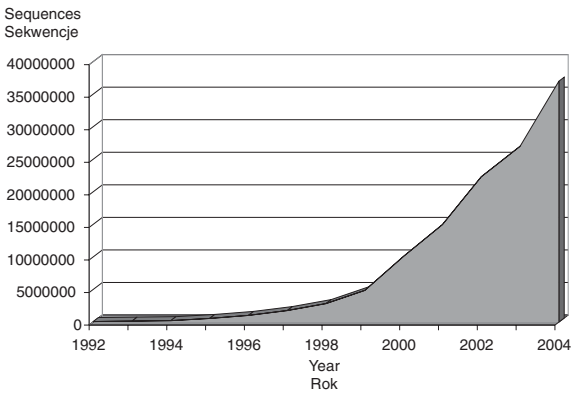

Fig. 3. Growth of GenBank during the period of 1992–2004

Ryc. 3. Wzrost liczby sekwencji w bazie GenBank w latach 1992–2004

SWISS-PROT

SWISS-PROT is protein-sequence database developed by the Department of Medical Biochemistry of the University of Geneva and the EMBL. Now it is a joint enterprise of the EMBL and the Swiss Institute of Bioinformatics (SIB). Sequence entries stored in SWISS-PROT are, when possible, formatted in the same way as those in the EMBL Nucleotide Sequence Database to maintain standards. Sequences stored in SWISS-PROT have two classes of data. Every sequence, apart from the core data (bibliographical references and a description of the source of the protein), has its own annotation, which is as accurate as possible. An annotation describes protein function, structure, post-translational modifications, domains and sites, similarities to other proteins, and sequence conflicts. All the required data are obtained from publications and review articles and checked by external experts [14].

The SWISS-PROT database minimizes the number of entries for a particular protein sequence in order to minimize redundancy. Each entry has a register of conflicts between sequencing reports. Entries are provided with links to other databases, so one can obtain as much information about the protein as possible, e.g. the nucleic acid sequence coding the particular protein, a description of genetic diseases, and the 3D structure of the protein. There are links to 29 different databases and the average entry has 3.5 cross-references. Many databases have been built based on SWISS-PROT to give the user a more precise view of a particular point of interest. The ProDom and DOMO databases contain a derived domain view of proteins. ProtoMap describes the classification of all SWISS-PROT proteins.

There are several model organisms on which SWISS-PROT focuses (chosen for genome sequencing and mapping projects). There are about 20 species, i.e. human, fruit fly, mouse, worm, and fission yeast. These organisms form the bulk of database entries, representing about 40% of all protein sequences. A large amount of data from several genome projects gave rise to TrEMBL (Translation of EMBL). TrEMBL contains entries derived from the translation of coding sequences (CDS) in the EMBL, but for CDS included in SWISS-PROT. TrEMBL has two sections: the first, SP-TrEMBL, stores sequences which await incorporation into SWISS-PROT, while the second, REM-TrEMBL, contains those not to be included, e.g. T-cell receptors, short amino-acid fragments, and sequences not coding a real protein [14].

The easiest way to access SWISS-PROT and TrEMBL is via the ExPASy (Expert Protein Analysis System) server at http://www.expasy.org. To acquire sequences, one can use Sequence Retrieval System (SRS) software. Another way of obtaining SWISS-PROT and TrEMBL is through anonymous ftp from ExPASy (ftp.expasy.org) or the European Bioinformatics Institute (EBI) (ftp.ebi.ac.uk/pub/) [14].

ProtoMap

Due to the increasing number of newly discovered sequences, the need arose to classify them and organize this large amount of data. Many sequences are of unknown function, so a common step is to compare such sequences with existing ones with already known properties, but in many cases the degree of divergence is so huge that a simple sequence comparison is unable to tell what the protein's function is. ProtoMap is a completely automatic method of protein-sequence classification. It searches for similarities by detecting groups of homologous proteins known as clusters and high-level structures, which are groups of related clusters. ProtoMap does not make use of multiple alignments. A two-phase algorithm is used to perform the analysis. The first step is an identification of groups of possibly related clusters. These (may) belong to the same group because they are strongly connected. After that, another test on a wider scale identifies a core of relationships in the group of clusters. The whole identification takes place at various confidence levels. Finally, an organization of all the proteins is made [15].

ProtoMap is strongly connected with SWISS-PROT and is based on the analysis of the database's protein sequences. The analysis made by ProtoMap is based on comparison of SWISS-PROT and TrEMBL, embracing almost all the known sequences. One can access ProtoMap at www.protomap.cs.huji.ac.il. There are several search methods available, from simple keywords through accession numbers to protein names. The essential matter is that we can classify a new protein sequence by submitting it and comparing it with existing clusters (it is possible that sequence will be ranked with several clusters in the order of connection quality) [15].

HOBACGEN and HOVERGEN

HOBACGEN (HOmologous BACterial GENes) is a database which contains all the available bacterial, archeabacterial, and yeast protein sequences from SWISS-PROT and TrEMBL. 109,077 protein sequences (78% bacteria, 13.8% archeabacteria, and 8.2% yeast) were stored in the database in the year 2000. But it is not only a collection of sequences. HOBACGEN offers multiple alignments and phylogenetic trees of homologous protein families. The main task of this database is to distinguish between orthologous and paralogous genes, which is achieved by a process of homology determination (aligning sequences and computing phylogenetic trees). Families were built using BLAST2 software; to measure amino-acid similarity, BLOSUM62 matrices were used. Multiple alignments were created under CLUSTAL_W, trees were computed in BIONJ. It is worth mentioning that the alignments did not undergo manual correction [16].

When using HOBACGEN, one can choose the desired gene family and select sequences homologous to those of particular interest. Then one has

several options to check: alignments, phylogeny, taxonomic data, and sequence annotations. All of them are information needed to distinguish orthologs from paralogs. These features form the main difference between HOBACGEN and other systems which allow checking homology relationships [16].

HOVERGEN (HOmologous VERtebrate GENe) is a project very similar to HOBACGEN. Instead of gathering bacterial sequences, it focuses on GenBank sequences from all vertebrate species. This database improves the quality of information which one can find in GenBank by clarifying several features (5' and 3' non-coding regions, introns, number of G-C at the 3rd position in codons) and correcting common errors. Similar to HOBACGEN, sequences were classified in families. These were aligned in CLUSTAL_V and phylogenetic trees were computed using the neighbor-joining method. Five families of the immune system were excluded from this process because of their complexity [17].

Both HOBACGEN and HOVERGEN focus on the same task. Their center of attention is minimizing redundancy. They allow users to obtain the necessary data with all the associated information and they offer a wide range of operations corresponding to selected sequences.

GOBASE

This database focuses on various data connected with mitochondria. GOBASE (http://megasun.bch.umontreal.ca/gobase/) contains sequences (both nucleotide and protein) encoded by mitochondrial genomes. All data undergo the process of standardization, thus ensuring that one has access to a complex site with sequences, RNA secondary structures, genetic maps, and taxonomic information. Moreover, new information is being added to the database all the time [18].

Organelles with their own genetic material (mitochondria and chloroplasts) are the subject of many studies. Especially mitochondria represent an interesting subject for scientists. Such intriguing processes as post-transcriptional RNA edition, oxidative phosphorylation, protein import, involvement in human diseases, and many others are all associated with these organelles. There is a wide range of information covering many fields of mitochondrial data, but all these particulars all scattered over the Internet and in various books and journals. Therefore it is a challenge for even an experienced researcher to find the piece of data he is interested in. The main idea of GOBASE is to integrate all the dispersed mitochondrial information and make it easily accessible through the Internet to a wide forum of users. Information stored in GOBASE is classified in ten categories, i.e. sequence, protein, taxon, etc. This organization is quite different from that used in the EMBL or NCBI. What is more, data in categories is cross-linked, consequently enabling more sophisticated queries. Apart from GenBank information, one can examine biochemical pathways of particular enzymes, RNA secondary structures, and mtDNA maps [18].

MIPS

MIPS stands for Munich Information Center for Protein Sequences. It incorporates several genome-specific databases and provides numerous tools for protein-sequence analysis. All sequences, obtained by various algorithms, are organized and associated to the proper information in scientific literature. The main projects drawn together under the wing of MIPS are the *Arabidopsis thaliana* genome (MATDB), Comprehensive Yeast Genome Database (CYGD), *Neurospora crassa* database (MNCDB), and the Database of Human cDNAs (DHGP). All of them share the same task: to classify and interpret complex data of genomes of interest. CYGD, for example, provides a set of useful information on functional associations between yeast genes and their proteins. One can choose from over 10,600 protein-protein interactions. There are also datasets on transcription factors (with binding sites), transport proteins, and metabolic pathways. SESAM, an incorporated tool, allows retrieving the related protein sequences from other species with high selectivity [19].

Molecular Data Analysis Tools

Gathering protein and gene data from molecular databases is just the beginning of bioinformatics work. The next step is to make proper use of the collected information. There are hundreds if not thousands of bioinformatics tools that can be used in every aspect of biological data examination, from phylogenetics to molecular modeling. Usually these tools are kindly provided and distributed by the authors. A short description of the tools used by the authors for phylogenetic analysis can be found below [20–22]. One should bear in mind that it was not this author's intention to describe all the available software, and there are different tools for sequence alignment, e.g. T-coffee, and phylogenetic analysis, e.g. PAUP and MOLPHY.

CLUSTAL_X

This is user-friendly software for multiple sequence alignment. The Clustal_X window is intuitive and easy to use, offering all the required options in pull-down menus. One can make use of several options to analyze aligned sequences. It is possible to highlight exceptional residues, change the order of sequences, and select a particular region of alignment to be realigned. Multiple alignment is usually built gradually. Closely associated groups are aligned first, then whole groups are aligned together. This method is not perfectly trustworthy, especially when sequences share less than 30% residue identity. The possibility to realign alignments prevails over some disadvantages of progressive alignment. One can deal with misaligned sequences in two ways. First it is possible to choose a sequence and realign it to another; the second option is to pick a particular range of the alignment. The program removes the selected region, makes a progressive alignment, and fits it back into the full arrangement. The quality of the alignment is checked by a score for each column in the alignment. A high score means that the column is highly conserved, while a low score indicates a less-well-conserved position. An alignment of loaded sequences is colored to highlight conservative regions. It might happen that highlighted residues are part of some significant area; such a situation is possible when a sequence has a new function compared with other sequences in a set. The user can define colors and set different colors depending on the level of residue similarity and/or physical properties [23].

SIFT

Sort Intolerant From Tolerant is a tool which can predict possible substitution sites in protein-coding regions. Because substitutions may have an influence on protein function, one is able to pick regions of higher priority to work on. Mutation can occur in regions of structural and functional importance, so predicting substitutions in these areas is of great significance. One of the great benefits of the program is formulating a prediction only from the sequence, not from the structure of the protein; thus it can identify a larger number of deleterious substitutions. SIFT uses a multiple alignment to check for deleterious and tolerated substitutions at every position of a sequence. Assuming that protein family members share highly conservative regions, the program searches for similar sequences, chooses the closest in function, and aligns them. After the alignment, the probability is calculated for every changeover at each position in the alignment. Substitutions with a probability less than a cutoff (0.05 by default) are predicted as deleterious, while those equal to or greater than the cutoff are predicted to be tolerated. For better results one should provide SIFT with a list of homologues rather than allow the program to choose them automatically. The most accurate predictions are made for orthologous sequences. As more and more protein sequences are available, their number will hopefully grow together with the constantly developing protein databases. The effects of SIFT predictions are much more accurate than the results of substitution scoring matrices. Furthermore, unlike matrices, SIFT does not need a lot of sequences to improve the prediction, offering a better result with only a single related one [24, 25].

MEGA3

Molecular Evolutionary Genetics Analysis software allows the user to perform various statistical and computational operations on amino-acid and nucleotide sequences. MEGA3 has a built-in Alignment Explorer, which includes CLUSTAL_W, as it is usually the first step in sequence comparison. Alignment can be done separately for coding and non-coding regions. The software implements a very useful option to put sequences into groups. During the analysis the user is free to operate on sequences. It is possible to exclude data with missing information (sites or codons) and alignment gaps. MEGA contains a number of methods for estimating the evolutionary distance between sequences. Distances are divided by the software into three factions: nucleotide, synonymous-nonsynonymous, and amino acid. Nucleotide distances approximate the number of nucleotide substitutions per site between DNA sequences. Synonymous and nonsynonymous distances are estimated by way of comparing codons between sequences using a specified genetic code table. Amino-acid sequences are translated from coding domains. For all amino-acid distances, methods are included to account for the rate of variation among amino-acid positions in the distance estimation. The program gives the user a powerful tool for constructing phylogenetic trees. It contains maximum parsimony (MP), minimum evolution (ME), neighbor-joining (NJ), and unweighted pair group with arithmetic mean (UPGMA) methods for inferring phylogenetic trees. It is worth mentioning that the current version of MEGA concatenates all the selected genes and domains, which makes inferring more accurate. After con-

struction, the tree is visualized in Tree Explorer, which can draw consensus and condensed trees. One can re-root the tree, change fonts and branch lengths, add scale bars, and more [26].

Conclusions

During the first years of bioinformatics, scientists developed a wide range of techniques useful not only in molecular evolution, but also for analyzing protein structure and functions. After modifications, these methods proved to be valuable in nucleid acid investigations. Although nowadays we may think of them as simple and uncomplicated, they still influence today's bioinformatics. Even now, the development of new tools and algorithms bears the mark of the pioneering work of the first bioinformaticians. Bioinformatic databases and tools have proved to be invaluable in the era of information overload. The variety of research topics connected with bioinformatics clearly shows the potential of computational analysis. With hundreds of thousands of sequences available, this discipline is rapidly growing, and the results of molecular genetic analysis are often impressive. Thanks to bioinformatics, a new field has opened for the investigation of the molecular basis of diseases and the evolutionary relationships between different organisms.

References

[1] Hagen JB: The origins of bioinformatics. Nature 2000, 1, 231–236.
[2] Sanger F, Thompson EO: The amino-acid sequence in the glycyl chain of insulin. Biochem. J. 1952, 52, iii.
[3] Doolittle RF, Singer SJ, Metzger H: Evolution of immunoglobulin polypeptyde chains: carboxy-terminal of an IgM heavy chain. Science 1966, 154, 1561–1562.
[4] Doolittle RF: The evolution of vertebrate fibrinogen. Fed. Proc. 1976, 35, 2145–2149.
[5] Miller CJ, Attwood TK: Bioinformatics goes back to the future. Nature 2003, 4, 157–162.
[6] Roberts RJ: The early days of bioinformatics publishing. Bioinformatics 2000, 16, 2–4.
[7] Trifonov EN: Earliest pages of bioinformatics. Bioinformatics 2000, 16, 5–9.
[8] Brown SM: Bioinformatics becomes respectable. BioTechniques 2003, 34, 2–5.
[9] Baxevanis AD: The molecular biology database collection: an online compilation of relevant database resources. Nucleic Acids Res. 2000, 28, 1–7.
[10] Piast M, Pałyga J: A diversity of chordate histone H1 complement. 12th International Symposium Molecular and physiological aspects of regulatory processes of the organism, Kraków, 2003, 317.
[11] Pałyga J, Piast M: Predicting tolerated amino acid substitutions in avian and mammalian somatic histone H1 variants. 12th International Symposium "Molecular and physiological aspects of regulatory processes of the organism, Kraków 2003, 300–301.
[12] Piast M, Kustrzeba-Wójcicka I, Banaś T: Molecular evolution of enolase. Acta Biochim Polon 2005, 52, 507–513.
[13] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: GenBank. Nucleic Acids Res 2004, 32, D23–D26.
[14] Bairoch A, Apweiler R: The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res 1999, 27, 49–54.
[15] Yona G, Linial L, Linial M: ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res 2000, 28, 49–55.
[16] Perriere G, Duret L, Gouy M: HOBACGEN: Database system for comparative genomics in bacteria. Gen Res 2000, 10, 379–385.
[17] Duret L, Mouchiroud D, Gouy M: HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res 1994, 22, 2360–2365.
[18] Korab-Laskowska M, Rioux P, Brossard N, Littlejohn TG, Gray MW, Franz Lang B, Burger G: The organelle genome database project (GOBASE). Nucleic Acids Res 1998, 26, 138–144.
[19] Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A: MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 2004, 32, D41–D44.
[20] Piast M, Kustrzeba-Wójcicka I, Banaś T: Enolase (EC 4.2.1.11) – a theoretical model of molecular evolution. Eur J Biochem 2004, 271, 78.
[21] Piast M, Kustrzeba-Wójcicka I, Banaś T: Functional and molecular diversity of enolase gene family. Eur J Biochem 2005, 272, 96.
[22] Piast M, Kustrzeba-Wójcicka I, Banaś T: Influence of evolution on molecular diversity of enolase – an enzyme of Embden-Meyerhof-Parnas pathway. Ukr Biokhim Zh 2005, 77(2), 137.
[23] 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, 25, 4876–4882.
[24] Ng CG, Henikoff S: Predicting deleterious amino acid substitutions. Gen Res 2001, 11, 863–874.

[25] Ng CG, Henikoff S: SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003, 31, 3812–3814.
[26] Kumar S, Tamura K, Nei M: MEGA3: Integrated software for Molecular Evolutionary Genetics analysis and sequence alignment. Brief Bioinform 2004, 5, 150–163.

Addres for correspondence:

Michał Piast
Department of Medical Biochemistry
Silesian Piast University of Medicine
Chałubińskiego 10
50-368 Wrocław
Poland
Phone: 071-7841381
piast@bioch.am.wroc.pl

Conflict of interest: None declared

Received: 26.04.2006
Revised: 11.05.2006
Accepted: 29.05.2006

Praca wpłynęła do Redakcji: 26.04.2006 r.
Po recenzji: 11.05.2006 r.
Zaakceptowano do druku: 29.05.2006 r.