ARTICLE
SYSTEM BIOLOGY & BIOINFORMATICS DOI: 10.5504/50YRTIMB.2011.0038

BIOLOGICAL SEQUENCE COMPARISON, MOLECULAR EVOLUTION AND PHYLOGENETICS

R.A. Dimitrov[1], D.E. Gouliamova[2]
[1]Sofia University of "St. Kliment Ochridski", Department of Physics, Sofia, Bulgaria
[2]Institute of Microbiology, Bulgarian Academy of Sciences, Sofia, Bulgaria
Correspondence to: Roumen Dimitrov
E-mail: roumen.dimitrov@gmail.com

ABSTRACT

For closely related sequences there is a single optimal alignment which provides an accurate measure of similarity, structure, function and evolutionary history. However, with increasing evolutionary distances between nucleotide sequences the single optimal alignment method is replaced by an ensemble of alignments of almost equal quality and ensemble of different self-folded conformations.

Recurring difficulties associated with diverged sequence data include alternative alignment possibilities of insertions and deletions, region of length variations in which homology assessment is questionable or impossible, occurrence of localized excessive mutations to the point of saturation and lost of phylogenetic signals. Therefore, for diverged sequences optimizing similarity will not necessarily improve structure, function and evolutionary history assessments.

Here our aim is to present an overview of the methods involved in sequence analysis which are critical for current theoretical and application development. However, we do not follow historical events. For sequence comparison we focus on those methods that are based on exhaustive schemes, which are classically formulated as dynamic programming algorithms. They consist either of optimization schemes which find the best alignment for a given model, or of probabilistic schemes based on partition functions - in which all alignments, with their respective weights, are evaluated.

Introduction

RNA/DNA and protein sequence data unite all organisms into the fold of comparative analyses allowing reconstruction of their evolutionary histories even they differ enormously in morphology and lifestyle. But while nucleotide and protein sequences are universal their tempo and mode of evolution are not.

Thus, mutation rates seem to vary both among and within genomes, being affected by many factors such as chromosomal position (16), G+C content (19), nearest neighbor bases (4), and different efficiency of the repair systems between the lagging and the leading DNA strands during replication and transcription (17).

On the other hand, sequences of biological macromolecules in various species, which share a common evolutionary ancestry, especially those with conserved catalytic activity, presumably fold into the same structure. Thus, for closely related species, optimizing similarity based on observed sequence variation can be used to obtain a single optimal alignment, which provides an accurate measure of similarity, structure, function and evolutionary history.

However, with increasing evolutionary distances between nucleotide sequences of distantly related species, the single optimal alignment method is replaced by an ensemble of alignments of almost equal quality and ensemble of different self-folded conformations.

Although the search for globally optimal similarity alignment is an ongoing process, the sequence alignment method diverged in its alignment objectives in a few major directions: 1) structure predictions (6); 2) database searching (3); 3) sequence comparison (8) and 4) phylogenetics (9).

The goal of structure prediction is to deduce the 2D and 3D structure of the gene product from a given gene sequence. The goal of alignment for database searching is to maximize the distinction between the homologous and non-homologous sequences. The major role of alignment for sequence comparison is to find out conserved sequence features (for example, functional sites). Finally, the goal of alignment for phylogeny is to align residues only if they have descended from common ancestral residue.

It is now evident that RNA/DNA and protein sequence evolution is far more complex than previously supposed and cannot be treated as an arbitrary string of characters, but is a macromolecule with specific biological constraints. The molecular structure and function may influence sequence evolution by generating sequence conservation or mutational hot spots of substitution or insertion/deletion events. Therefore, sequence alignment should model molecular processes that have led to the observed sequence variation rather than similarity-based patterns.

Nucleotide substitution models

We call two biological sequences homologous if they are evolutionarily similar, i.e., were derived from a common ancestor. For any two sequences we want to be able to compare them and to see whether they are homologous or not and to measure the extent of this homology. But first we need a model for nucleotide and amino acid substitutions which cause the sequence change in the course of time.

Substitutions are usually modeled as a random event. The simplest approach to account for a random mutation at a particular site along the sequence is based on the application of the Poisson process in probability theory. This approach was used by Zuckerkandl and Pauling (20) in predicting the evolutionary change of hemoglobin and cytochrome c. Zuckerkandl and Pauling proposed the theory of a molecular clock, that is, that the rate of molecular evolution is approximately constant over time for all the proteins in all lineages (**Fig. 1**).

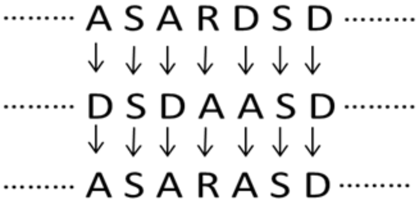


Fig. 1. The evolutionary process operates independently on each site of the protein sequences. The arrows represent substitution events during time. On each site of the sequences all nucleotide substitutions occur at an equal rate, and when a nucleotide is substituted, any one of the other nucleotides is equally likely to be its replacement

According to this theory, any time of divergence between genes, proteins, or lineages can be dated simply by measuring the number of changes between sequences. Here we will shortly follow their considerations. Let λ be the rate (probability) of amino acid substitution per year at a particular amino acid site and assume that it remains constant for the entire evolutionary period. Then the mean number of amino acid substitutions at this site during a period of t years is λt, and the probability of occurrence of r amino acid substitutions is given by $P_r(t) = e^{-\lambda t}(\lambda t)^r/r!$.

Since the probability that amino acid substitution does not occur at a particular site during t years is $e^{-\lambda t}$, the probability that neither of the homologous sites of the two sequences from a pair of species undergoes substitution is $e^{-2\lambda t}$. If λ is the same for all amino acid sites, the expected number of identical amino acids n_i between the two sequences is: $n_i = ne^{-2\lambda t}$

Therefore, the simplest approach to measure the divergence between two closely related strands $\left(1 \gg \frac{n-n_i}{n}\right)$ of aligned sequences is to count the number of sites where they differ $(n - n_i)$. Thus, evolutionary time of divergence $T = 2\lambda t$ can be estimated from:

$$T = 2\lambda t = -log(1-p) = -ln\left(1 - \frac{n-n_i}{n}\right) \sim \frac{n-n_i}{n}$$

The quantity $p = \frac{n-n_i}{n}$ is called p-distance. Unfortunately, if the rate of substitution is high the p-distance is generally not very informative with regard to the number of substitutions that actually occurred. This is due to the fact that the formula does not include the possibility that two or more mutations can take place consecutively at the same site and the case of a back-mutation. Therefore, the p-distance is reasonable only for closely related sequence.

It is difficult to implement, in a model that aims to be general, all the different mutation rules and patterns that we detect in the genetic material belonging to different species. Therefore, we will continue our discussion about substitution models with a description of an assumption they all share, the Markov property.

The approach outlined in the Poisson process can be generalized to a so-called Markov process (8). Consider a stochastic model for RNA/DNA or amino acid sequence evolution. We assume independence of evolution at different sequence sites and thus can consider sites one by one. At any single site, the model works with probabilities $P_{ij}(t)$ that base i will have changed to base j after a time t. The subscripts i and j take the values 1,...,4 to represent the nucleotides A,C,G, and $T(U)$ for RNA/DNA sequences and 1,...,20 for amino acid sequences.

The Markov process uses a Q matrix that specifies the relative rates of change of each nucleotide or amino acid along the sequence. In the case of RNA/DNA sequences rows and columns of the Q matrix follow the order A,C,G, and $T(U)$, so that, for example, the third term of the second row is the instantaneous rate of change from nucleotide C to nucleotide G. The most general independent site evolution rate matrix has twelve different parameters as shown in **Fig. 2**:

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \left(\begin{array}{cccc} - & Q_{AC} & Q_{AG} & Q_{AT} \\ Q_{CA} & - & Q_{CG} & Q_{CT} \\ Q_{GA} & Q_{GC} & - & Q_{GT} \\ Q_{TA} & Q_{TC} & Q_{TG} & - \end{array}\right) \end{array} \quad - = -\sum_{i \neq j} Q_{ij}$$

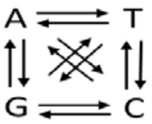


Fig. 2. Site evolution rate matrix Q for nucleotides. Each site in the RNA/DNA sequence is treated as a random variable with a discrete number n of possible states. For nucleotides there are four states ($n = 4$) which correspond to the four nucleotide bases A,C,G, and $T(U)$. The components of the rate matrix Q correspond to the rate of replacement of one nucleotide by another. In other words, each non-diagonal entry in the matrix represents the flow from nucleotide i to j, while the diagonal elements are chosen in order to make the sum of each row equal to zero since they represent the total flow that leaves nucleotide i

The stochastic process for substitution events can be derived from first principles such as detailed balance and the Chapman-Kolmogorov equations (5). To model the substitution process on the RNA/DNA level for example it is commonly assumed that a replacement of one nucleotide by another occurs randomly and independently, and that nucleotide frequencies π_i in the data do not change over time and from sequence to sequence in an alignment. In other words, we can say that the dynamical properties of the stochastic substitution process do not vary in time and therefore, π_i is the equilibrium distribution that the stochastic substitution process reaches asymptotically. The equilibrium distribution has the property of being a fixed point of the dynamic stochastic process. It can thus be obtained by solving the eigenvalue problems:

$$\pi_j = \sum_k P_{jk}(t)\,\pi_k \quad \text{or} \quad 0 = \sum_k Q_{jk}\,\pi_k$$

The transition probabilities can be represented in the form of transition probability matrix $P(t)$. It's components $P_{ij}(t)$ satisfy the conditions:

$$\sum_j P_{ij}(t) = 1 \text{ and } P_{ij}(t) > 0 \text{ for } t > 0$$

The probability matrix $P(t)$ also satisfies the Chapman-Kolmogorov equation:

$$P(t+s) = P(t)\,P(s)$$

and the initial conditions $P_{ij}(0) = 1$, for $i = j$ and $P_{ij}(0) = 0$ for $i \neq j$. It is also often assumed that the substitution process is reversible:

$\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$ and $Q_{ij} = R_{ij}\pi_j$ for $i \neq j$, where $R_{ij} = R_{ji}$ is a symmetric matrix of rate parameters with vanishing diagonal elements $R_{ii} = 0$. Any model in use today follow from a particular choice of nucleotide substitution in the framework of a reversible rate matrix by specifying explicit values for the matrix R and for the frequencies π_i.

The transition rate matrix Q is a first order approximation for the Markov process. From it we can derive the basic equation that describes the dynamic of a Markov process Thus, for small t keeping only the linear terms the transition probability matrix $P(t)$ can be represented in Taylor expansion in the form:

$$P(t) = P(0) + tQ \quad \text{or} \quad Q = \lim_{t\to 0}\frac{P(t)-I}{t}$$

where, I is the identity matrix. This equation provides an infinitesimal description of the substitution process. From the Chapman-Kolmogorov equation we get the forward and backward differential equations:

$$\frac{d}{dt}P(t) = \lim_{t\to 0}\frac{P(t+dt)-P(t)}{t} = \lim_{t\to 0}\frac{P(t)-I}{t}P(t) = QP(t) = P(t)Q$$

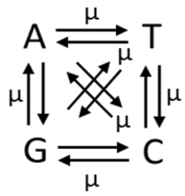
The solution of these equations, taking into account the initial condition is:

$$P(t) = \exp(tQ) = \sum_n^{\infty} Q^n \frac{t^n}{n!}$$

Now we can compare the results for evolutionary divergence between two sequences in the Markov process and the Poisson process. The total number of substitutions per unit time, i.e. the total rate μ is $\mu = -\sum_i \pi_i Q_{ii}$ therefore, the number of substitutions during time t is $T = t\mu = -t\sum_i \pi_i Q_{ii}$. The probability that a substitution is observed after time t is $p = 1 - \sum_i \pi_i P_{ii}(t)$.

The simplest case of Markov process is the Jukes-Cantor model (8). In this model we have the following assumptions: First, each nucleotide position in the sequence evolves independently from all others; Second, transition probabilities are all equal to the same value. Each nucleotide is equally likely to turn into each other nucleotide. Therefore, for the Jukes-Cantor model we have:

$$Q = \begin{pmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{pmatrix}$$



$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}exp(-4\mu t)$$
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}exp(-4\mu t)$$

Thus, for two sequences that split t time units ago we have $p = \frac{3}{4}\left(1 - exp(-8\mu t)\right)$. Finally, for the evolutionary time of divergence or the genetic distance between two sequences we have:

$T = 2(3\mu t) = -\frac{3}{4}log\left(1 - \frac{4}{3}p\right)$ Jukes-Cantor model

$T = 2\lambda t = -log(1-p)$ Poisson model

The theory of Markov processes provides us with a powerful formalism to describe the evolution in time of a single genomic sequence. However, we have no direct observations of how species or genomes evolve. What we have are instead sets of genomic sequences of different present day species.

In the past, species evolution and relationships was inferred by examining fossil records and morphological characters. In the 1960's, when molecular techniques were introduced to the field, the evolutions of the organisms' macro-molecules were used to reconstruct their evolutionary history (14). This approach is based on the assumption that sequences from different species have descended from some ancestral gene in a common ancestral species. Thus, divergence between sequences is a result of speciation. These genes are essentially the same and are called orthologues. For some sequences the assumption may not hold because of gene duplication, another mode of evolution.

Such relationship between species is called *phylogeny* (8, 9). The simplest approach which can translate mathematically the concepts of common descent is to assume that the evolution of species or genomes can be presented as a Markov process on a tree like structure called *phylogenetic tree* (**Fig. 3**).

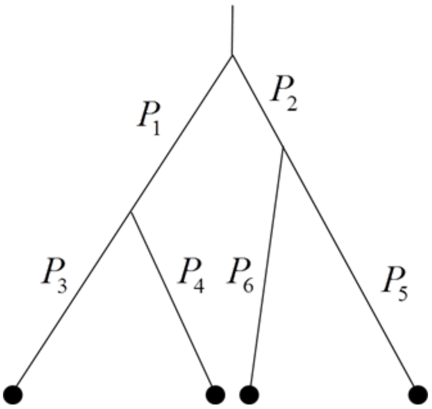


Fig. 3. An example of phylogenetic Markov process
The tree represents 4 present day species, and 3 extinct ones, the common ancestors. Along its branches there are 6 different Markov processes representing the sequence evolutionary dynamics

Phylogenetic trees are often used to present the history of molecules (14). The root of the tree is regarded as the common ancestor of all the sequences. Internal nodes in the tree represent divergence points. The length of each edge of the tree represents the amount of evolutionary divergence between sequences. These lengths do not necessarily correspond to evolutionary time periods, because sequences evolve at different rates in different organisms.

If the exact amount of sequence divergence between all pairs of sequences from a set of n sequences is known, the genetic distance between the sequences provides a basis to infer the evolutionary tree relating the sequences (8, 9).

Sequence Comparison

Before the similarity or the genetic distance of two sequences can be evaluated, one typically begins by finding a plausible alignment between them. A sequence alignment is simply an array where each row corresponds to one of the sequences and where those bases which are assumed to be homologous to each other stand in the same column. We get such an alignment by inserting special characters $("-")$ called gaps which describe the insertion and deletion events. In the case that our alignment covers the entire sequences we speak of a global alignment. If we are only making assumptions about the relatedness of some parts of the sequences we call it local alignment. An alignment which only has two rows is called a pairwise alignment. If it has more than two rows we call it a multiple alignment (**Fig. 4**).

$$\mathcal{A}^{\text{global}} = \begin{pmatrix} \mathrm{A-TAC-TGG} \\ \mathrm{-GTCCGT-G} \end{pmatrix}$$

$$\mathcal{A}^{\text{local}} = \begin{pmatrix} \mathrm{A-TAC-TGG} \\ \mathrm{TCCGT} \end{pmatrix}$$

Fig. 4. Pairwise global $\mathcal{A}^{\text{global}}$ and local $\mathcal{A}^{\text{local}}$ alignments

For example, in the $\mathcal{A}^{\text{global}}$ alignment in **Fig. 4** the two sequences are identical in the third, fifth, seventh, and ninth columns – they are the matches. There is a mismatch in the fourth column, first and eighth columns are deletions, and second and sixth columns are insertions.

The concept of an alignment is crucial (15). In the case of biological sequence comparison we want a biologically relevant alignment. This can be achieved by using some essential molecular evolution assumptions combined with probabilistic techniques (8). An alignment should represent a specific hypothesis about the evolution of the sequences and its purpose is to find the alignment which maximizes the probability of two sequences having evolved from a common ancestor as opposed to being just random sequences. We do this by having models that assign a probability to the alignment in each of the two cases and then consider the ratio of the two probabilities.

Consider a pair of sequences A and B, of lengths N and M, respectively. Let A_i be the ith symbol of A and B_j be the jth symbol of B. These symbols will come from some alphabet $\mathcal{A}$. In the case of DNA $\mathcal{A}$ ={A, T, G, C} and in the case of proteins the twenty amino acids. Given a pair of aligned sequences, we want to assign a score to the alignment that gives a measure of the relative likelihood that the sequences are related as opposed to being unrelated (**Fig. 5**).

Let (as shown in **Fig. 5**) $p_{A_iB_j}$ be the probability that the nucleotides A_i and B_j have each independently derived from the same original residue in their common ancestor, while p_{A_i} and p_{B_j} are the probabilities of occurrence of A_i and B_j in their sequences if we consider them as random. We want the score for aligning nucleotides A_i and B_j $s(A_i, B_j)$ to be the log likelihood ratio of nucleotide pair (A_i, B_j) occurring as an aligned pair, as opposed to a nonaligned pair. Therefore, we have:

$$s(A_i, B_j) = klog\left(\frac{p_{A_iB_j}}{p_{A_i}p_{B_j}}\right)$$

It is intuitively obvious that taking random sequences that were generated according to a given underlying letter-distribution, and aligning these sequences randomly, the frequency of letter pairs should differ depending on the distribution of the letters. In aligning biological sequences one would expect identities and conservative substitutions to be more frequent than they would appear in chance alignments $(p_{A_iB_j} > p_{A_i}p_{B_j})$. Therefore, these columns should contribute positive terms to the score of an alignment $(s(A_i, B_j) > 0)$. Similarly, non-conservative changes should be less frequent, than they would be in chance alignments $(p_{A_iB_j} < p_{A_i}p_{B_j})$. Therefore, the columns corresponding to non-conservative changes should contribute negative terms to the alignment score $(s(A_i, B_j) < 0)$.

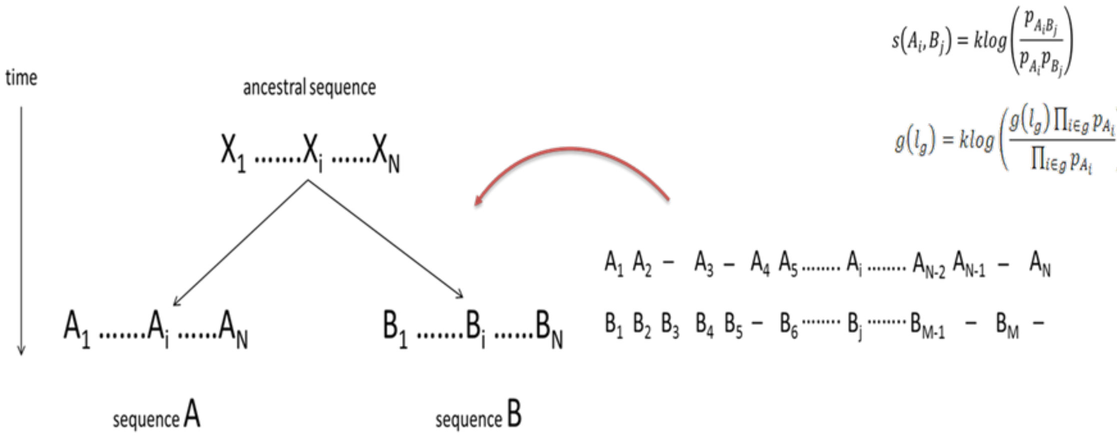

$s(A_i,B_j) = klog\left(\frac{p_{A_iB_j}}{p_{A_i}p_{B_j}}\right)$
$g(l_g) = klog\left(\frac{g(l_g)\prod_{i\in g} p_{A_i}}{\prod_{i\in g} p_{A_i}}\right)$
time
ancestral sequence
X1 ....Xi ......XN
A1 ....Ai ......AN
B1 ....Bi ......BN
sequence A
sequence B
A1 A2 – A3 – A4 A5........ Ai .... AN-2 AN-1 – AN
B1 B2 B3 B4 B5 – B6........ Bj .... BM-1 – BM –

Fig. 5. Alignment represents a specific hypothesis about the evolution of the sequences

The question is how to determine the parameters in the scoring model. There are two main approaches: 1) counting the frequencies of aligned pairs and of gaps in confirmed alignments, and to set the probabilities $p_{A_iB_j}$, p_{A_i}, p_{B_j} and $g(l_g)$ to the normalized frequencies. In this approach one have to be careful in assembling a random sample of confirmed aligned sequences, because for example protein sequences come in families; 2) if sequences evolve on a given phylogenetic tree, knowing the evolutionary rate matrix Q and the stationary distribution of the sequences we can estimate $p_{A_iB_j}$ and therefore, the score for aligning nucleotides A_i and B_j has the form:

$$s(A_i, B_j|t) = klog\left(\frac{e^{tQ_{A_iB_j}}}{p_{B_j}}\right)$$

where t is the degree of evolutionary divergence that we are focusing on and $e^{tQ_{A_iB_j}} = \frac{p_{A_iB_j}}{p_{A_i}}$ is the conditional probability that A_i is replaced by B_j in time t. The above expression follows from the stationarity and time-reversibility of the Markov process (8). In this approach one have to be careful with the fact that different pairs of sequences have diverged by different amounts. Thus, when two sequences have diverged from a common ancestor very recently, many pairs will be identical. Therefore, $p_{A_iB_j} \ll 1$ for $A_i \neq B_j$ and $s(A_i, B_j) < 0$. In the opposite case $p_{A_iB_j} \approx p_{A_i}p_{B_j}$ and $s(A_i, B_j) \approx 0$.

Specifying an appropriate score matrix is central to sequence comparison methods, and much effort has been devoted to defining, analyzing, and refining such matrices (2, 10, 11, 15).

The probability of a gap occurring at a particular site in the sequence is the product of a function $g(l_g)$ of the gap-length and the combined probability of the set of inserted nucleotides $\prod_{i\in g} p_{A_i}$ (1, 8).

$$P(g) = g(l_g)\prod_{i\in g} p_{A_i}$$

The total score of the alignment $\boldsymbol{\mathcal{A}}$ of sequence A and sequence B is sum of scores for each aligned pair of nucleotides, plus the term for gaps.

$$S(A,B) = \sum_{(ij)\in\boldsymbol{\mathcal{A}}} klog\left(\frac{p_{A_iB_j}}{p_{A_i}p_{B_j}}\right) + \sum_{l_g} klog\left(g(l_g)\right)$$

Dynamic programming algorithms

Because of the time-reversibility of the substitution model - there are not any restrictions on the Markov process that operates at the variable sites other than that it is stationary and reversible - the likelihood that one sequence evolved into the other is twice the time that separates the ancestor from the two descendants. This likelihood is, by definition, the total probability corresponding to all evolutionary histories that are consistent with the observed sequences. Waterman (18) presents a brief combinatorial treatment of alignments to estimate the number of different alignments between two sequences with lengths N and M. He points out that If one does not count permutations such as:

$$\begin{matrix}\mathrm{G}-\\-\mathrm{C}\end{matrix} \quad \text{and} \quad \begin{matrix}-\mathrm{G}\\\mathrm{C}-\end{matrix}$$

there are $\binom{M}{k}\binom{N}{k}$ alignments with k aligned pairs. Therefore, all alignments are:

$\sum_{k\geq 0}\binom{M}{k}\binom{N}{k} = \binom{M+N}{N}$ and for $N = M$ we have $\binom{2N}{N} \approx \frac{1}{\sqrt{N\pi}}2^{2N}$

Thus, for two sequences of length 1000 we have $\frac{1}{\sqrt{1000\pi}}2^{2000} \approx 10^{600}$ alignments. Obviously, there are extremely many of these evolutionary histories, so a direct evaluation of this sum is impractical. However, a dynamic programming approach is possible that computes this sum in reasonable time.

Here we examine the exhaustive schemes, which are classically formulated as dynamic programming algorithms. They consist either of optimization schemes which find the best alignment for a given model, or of probabilistic schemes based on partition functions - in which all alignments, with their respective weights, are evaluated.

The first method for generating sequence alignments based on Dynamic programming was described by Needleman and Wunsch (13) and was based on maximizing the similarity score between sequences. The idea of dynamic programming is to build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences (8).

Consider a pair of sequences A and B, of lengths N and M, respectively. The goal is to compute $S(N,M)$. To achieve it we need to evaluate the best intermediate alignments $S(i,j)$ for the substrings $(A_1, A_2, \dots A_i)$ and $(B_1, B_2, \dots B_j)$ as shown in **Fig. 6**.

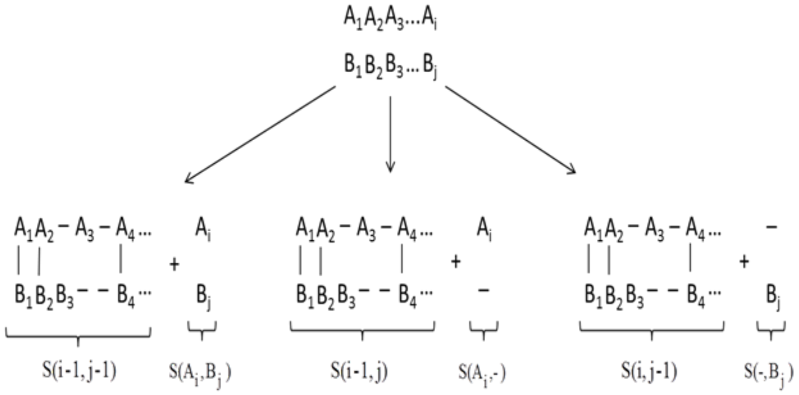

A1A2A3...Ai
B1B2B3...Bj
A1A2 – A3 – A4... Ai + B1B2B3– – B4... Bj
A1A2 – A3 – A4... Ai + B1B2B3– – B4... –
A1A2 – A3 – A4... – + B1B2B3– – B4... Bj
S(i-1,j-1)
S(Ai,Bj)
S(i-1,j)
S(Ai,-)
S(i,j-1)
S(-,Bj)

Fig. 6. To express the score $S(i,j)$ we decompose partial alignments into three cases that can occur at the ends A_i, B_j: 1) A_i matches to B_j - $\binom{A_i}{B_j}$; 2) A_i is alignment to gap - $\binom{A_i}{-}$ and 3) B_j is alignment to gap - $\binom{-}{B_j}$. Additionally we need to know values of $S(i-1,j)$, $S(i,j-1)$, $S(i-1,j-1)$.

We start by initializing:

$$\begin{cases} S(0,0) = 0 & \text{the score for alignment at the beginning} \\ S(i,0) = klog\big(g(i)\big) & \text{the score for alignment of all residues of sequence A to gaps} \\ S(0,j) = klog\big(g(j)\big) & \text{the score for alignment of all residues of sequence B to gaps} \end{cases}$$

This process can be expressed more formally:

$$S(i,j) = max\begin{cases} S(i-1,j-1) + S\big(A_i,B_j\big) & corresponds\ to\ \binom{A_i}{B_j} \\ S(i,-1j) + S(A_i,-) & corresponds\ t\text{□}\ \binom{A_i}{-} \\ S(i,j-1) + S\big(-,B_j\big) & corresponds\ to\ \binom{-}{B_j} \end{cases}$$

The above equation presents a recursive relation between adjacent cells in the array S. In that manner we compute iteratively consecutive values of cells in the S matrix obtaining a path from top left to the bottom right. It is important to note that all possible alignments between two sequences correspond one-to-one to such directed paths in the S matrix. The best score for an alignment of sequences $(A_1, A_2, \ldots A_N)$ and $(B_1, B_2, \ldots B_M)$, by definition, is the value of the final cell of the matrix S (**Fig. 7**).

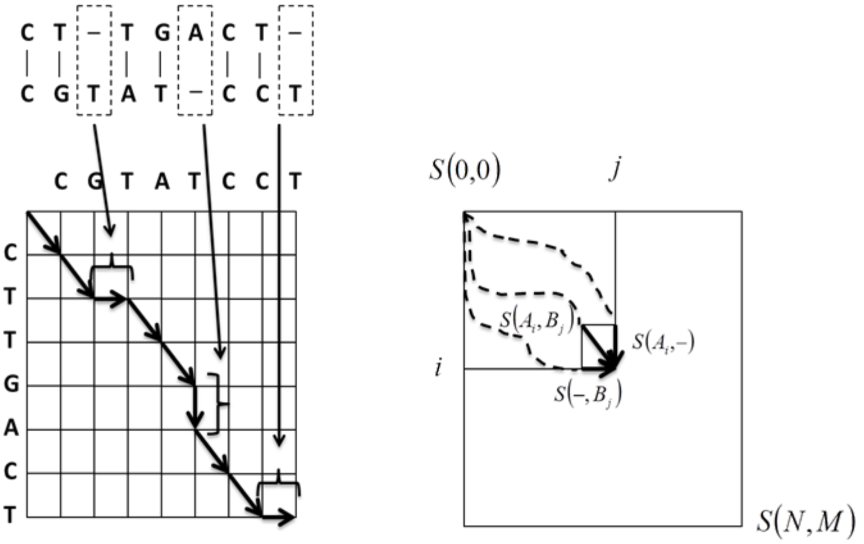


Fig. 7. Example of global alignment path for sequences CGTATCCT and CTTGACT, and path iterative cell by cell extension

To find the path we need to trace back through the cells with maximum value of $S(i,j)$. Local sequence alignment problem can be easily solved by modification of the algorithm for the global alignment. Thus, when adjacent cells have a negative score then we have to assign 0 score for the considered cell.

By using the approach of path representation of suboptimal alignments, one can miss alignments with biological correct solutions. This can be avoid by taking into account the partition function of all possible paths or alignments between two sequences. Computation of the partition function for alignments was pioneered by Miyazawa (12).

The score for the alignment of two sequences A and B is the sum of the scores for all gaps in the alignment, plus the sum of the scores for all substitutions:

$$S(\boldsymbol{\mathcal{A}}) = k\left\{\sum_{(ij)\in\boldsymbol{\mathcal{A}}} log\left(\frac{p_{A_iB_j}}{p_{A_i}p_{B_j}}\right) + \sum_{l_g} log\left(g(l_g)\right)\right\}$$

Therefore, for the probability of a particular alignment between the sequences A and B we can write:

$$e^{S(\boldsymbol{\mathcal{A}})} = e^{k\left\{\sum_{(ij)\in\boldsymbol{\mathcal{A}}} log\left(\frac{p_{A_iB_j}}{p_{A_i}p_{B_j}}\right) + \sum_{l_g} log\left(g(l_g)\right)\right\}} = \left(\frac{\prod_{l_g} g(l_g) \prod_{(ij)\in\boldsymbol{\mathcal{A}}} p_{A_iB_j}}{\prod_{(ij)\in\boldsymbol{\mathcal{A}}} p_{A_i}p_{B_j}}\right)^k = \left(\frac{P(\boldsymbol{\mathcal{A}})}{P(A)P(B)}\right)^k$$

where $P(A) = \prod_i p_{A_i}$ and $P(B) = \prod_j p_{B_j}$ are the random probabilities for the sequences A and B. Therefore, $P(\boldsymbol{\mathcal{A}}) \approx e^{\frac{S(\boldsymbol{\mathcal{A}})}{k}}$

The sum over the probabilities of all possible alignments $\boldsymbol{\mathcal{A}}$ between the two sequences A and B has to be 1:

$\sum_{\boldsymbol{\mathcal{A}}} P(\boldsymbol{\mathcal{A}}) = c \sum_{\boldsymbol{\mathcal{A}}} e^{\frac{S(\boldsymbol{\mathcal{A}})}{k}} = 1$ and therefore, $P(\boldsymbol{\mathcal{A}}) = \frac{e^{\frac{S(\boldsymbol{\mathcal{A}})}{k}}}{\sum_{\boldsymbol{\mathcal{A}}} e^{\frac{S(\boldsymbol{\mathcal{A}})}{k}}}$

With analogy to statistical physics we can introduced a partition function:

$$P(\boldsymbol{\mathcal{A}}, \mathrm{T}) = \frac{e^{\frac{S(\boldsymbol{\mathcal{A}})}{k}}}{Z(T)} \text{ and } Z = \sum_{\boldsymbol{\mathcal{A}}} e^{\frac{S(\boldsymbol{\mathcal{A}})}{kT}}$$

where the parameter T plays the role of temperature. The partition function is computed by means of a dynamic programming algorithm and used to determine the probability of an alignment as well as the probability of each possible match between two sequence positions. Using the additive property of the alignment score:

$$S(\boldsymbol{\mathcal{A}}) = S^{forward}(i,j) + S^{backward}(i,j) - S(A_i, B_j)$$

We can write for the match probability (7):

$$P(A_i, B_j) = \frac{Z^{forward}(A_i, B_j) Z^{backward}(A_i, B_j)}{Ze^{\frac{S(A_i,B_j)}{kT}}}$$

where, $Z^{forward}(A_i, B_j)$ and $Z^{backward}(A_i, B_j)$ are calculated in exactly the same way as $S(i,j)^{forward}$ and $S(i,j)^{backward}$ with exception that instead taking the max we are summing.

Conclusions

In this article our aim was to present an overview of the methods involved in sequence analysis from computational, biological, and statistical perspectives without following closely the historical events. We focus on those of them which are critical for current theoretical and application development.

Thus, the acquisition of large multilocus sequence data is providing researchers with unprecedented amount of information. With these large quantities of data comes the increasing challenge regarding the best methods of sequence analysis.

The comparison of a new sequence against a database of known sequences is perhaps the most important computer application in molecular sequence analysis. For closely related sequences there is a single optimal alignment which provides an accurate measure of similarity, structure, function and evolutionary history. However, with increasing evolutionary distances between nucleotide sequences the single optimal alignment method is replaced by an ensemble of alignments of almost equal. Moreover, with increasing evolutionary distance single optimal alignment methods tend to become sensitive to the choice of scoring parameters and therefore less reliable. In this regard we have also briefly focused on methods involved in determination of scoring parameters.

The molecular structure and function may influence sequence evolution by generating sequence conservation or mutational hot spots of substitution or insertion/deletion events. Therefore, the probability of substitutions and the overall mutation rate varies in different regions of biological sequences. In this regard, knowledge of what parts of an alignment are reliably aligned and what parts display a high degree of ambiguity is of increasingly importance.

It is generally accepted that the Smith-Waterman local similarity search algorithm is the most sensitive technique to discover significant weak similarities between two sequences. However, this approach is based on a limited sample of suboptimal alignments and one can miss alignments with biological correct solutions. Therefore, we present the theory of probabilistic alignment derived from a thermodynamic partition function which can avoid this problem taking into consideration all possible alignments between two sequences. A probabilistic notion of alignment is also inherent to information-theoretic approaches.

Recently there is increased interest in regards to thermodynamic partition function from computational methods for predicting evolutionarily conserved rather than thermodynamic structures of RNA and protein molecules. Recurring difficulties associated with alignment of structurally related, but otherwise diverged sequences include alternative alignment possibilities of insertions and deletions, region of length variations in which homology assessment is questionable or impossible, occurrence of localized excessive mutations to the point of saturation and loss of phylogenetic signals. Therefore, for diverged sequences optimizing similarity will not necessarily improve structure, function and evolutionary history assessments.

Development of a method based on thermodynamic partition function for structurally related, but diverged sequences for simultaneous optimization of alignment and self-folding - the so-called Sankoff's program for simultaneous prediction of secondary structure and alignment between nucleotide sequences - is great challenge. Still there is not a general solution for this long standing problem.

REFERENCES

1. **Altschul S.F. and Erickson B.W.** (1986) Bull. Math. Biol., **48**(5-6), 603-616.
2. **Altschul S.F.** (1991) J. Mol. Biol., **219**(3), 555-565.
3. **Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.** (1997) Nucleic Acids Research, **25**(17), 3389-3402.
4. **Blake R.D., Samuel T.H., Nicholson-Tuell J.** (1992) J. Mol. Evol., **34**, 189-200.
5. **Blossey R.** (2006) Computational Biology: A Statistical Mechanics Perspective, Chapman & Hall/CRC.
6. **Christian L. and Tamar S.** (2011) Current Opinion in Structural Biology, **21**, 1-13.
7. **Dimitrov R.A.** (2005) Bulg. J. Phys., **32**, 220-235.
8. **Durbin R., Eddy S., Krogh A., Mitchison G.** (1998) Biological sequence analysis, Probabilistic models of proteins and nucleic acids, Cambridge University Press.
9. **Felsenstein J.** (2004) Inferring Phylogenies, Sunderland MA, Sinauer Associates.
10. **Henikoff S. and Henikoff J.G.** (1992) Proc. Natl. Acad. Sci. USA, **89**(22), 10915-10919.
11. **Henikoff S.** (1996) Curr. Opin. Struct. Biol., **6**(3), 353-360.
12. **Miyazawa S.** (1994) Protein Eng., **8**(10), 999-1009.
13. **Needleman S.B. and Wunsch C.D.** (1970) J. Mol. Biol., **48**(3), 443-453.
14. **Olsen G.J. and Woese C.R.** (1993) FASEB, **7**, 113-123.
15. **Rosenberg M.S.** (2009) Sequence alignment: methods, models, concepts, and strategies, University of California Press, Ltd, Berkeley, CA.

16. **Sharp P.M., Shields D.C., Wolfe K.H., Li W.-H.** (1989) Science, **258**, 808-810.
17. **Veaute X. and Fuchs R.** (1993) Science, **261**, 598-601.
18. **Waterman M.S.** (1995) Introduction to Computational Biology: Maps, Sequences and Genomes, Chapman & Hall/CRC.
19. **Wolfe K.H.** (1991) J. Theor. Biol., **149**, 441-451.
20. **Zuckerkandl E. and Pauling L.** (1965) In: Evolving genes and proteins (V. Bryson and H. Vogel, Eds.), Academic Press, New York, NY, 97-166.