ARTICLE
SYSTEM BIOLOGY & BIOINFORMATICS **DOI: 10.5504/50YRTIMB.2011.0039**

NEW METHOD FOR SEQUENCE ALIGNMENT BASED ON PROBABILITIES OF NUCLEOTIDE CORRESPONDENCES

R. Dimitrov[1], D. Gouliamova[2]
[1]University of "Saint Kliment Ohridski", Sofia, Bulgaria
[2]Institute of Microbiology Bulgarian Academy of Sciences, Sofia, Bulgaria
Correspondence to: Roumen Dimitrov, Dilnora Gouliamova
E-mail: roumen.dimitrov@gmail.com; dilnorag@gmail.com

ABSTRACT

The objective of our work is to develop a general method 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. A simple reason behind the simultaneous optimization of alignment and self-folding is that strong structural consensus among related, but diverged sequences are a good indicator for preserved functional role. Up to now there is no a general solution for this long standing problem.

Here we discuss an approach which is just a first step to the full realization of Sankoff's program. Currently available models and software packages, such as foldalign, dynalign and others, implement only restricted versions (variations around first align and then fold or oppositely) of Sunkoff's program and do not use the full loop-based RNA/DNA energy model.

We divided Sankof's program in two steps based on the analogy between the classical alignment algorithm and hybridization without self-folding. The next step is to include in the alignment an algorithm for the self-folding.

In our approach, the alignment problem requires the implementation of the full loop-based RNA/DNA energy model for hybridization of two sequences. For this, we divided the alignment between two sequences into loops and associated a score to each loop in such way that the total score of the alignment is a sum over the scores for each alignment loop. The loop scoring model for alignment consists of following loop types: stacking with matched and mismatched pairs, bulges, internal loops and dangling ends.

Calculation of thermodynamic partition function over all possible double-stranded conformations is interpreted in terms of all possible canonical pairwise alignments. 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. For calculation of match probabilities detailed recursion relations for partition functions of alignments are based on their recursion analogs for hybridization of subsequences. The partition function is used for backtracking and reconstructing a properly weighted ensemble of optimal and suboptimal alignments.

Introduction

DNA sequence data unite all organisms into the fold of comparative analyses allowing reconstructing their evolutionary histories even they differ enormously in morphology and lifestyle (7). But while nucleotide sequences are universal their tempo and mode of evolution are not (8, 17).

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 (3, 11, 14).

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 (2, 10, 20).

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 (7, 12). Therefore, for diverged sequences optimizing similarity will not necessarily improve structure, function and evolutionary history assessments.

The objective of our work is to develop a general method for structurally related, but diverged sequences for simultaneous optimization of alignment and self-folding - the so-called Sankoff's program (18) for simultaneous prediction of secondary structure and alignment between nucleotide sequences. A simple reason behind the simultaneous optimization of alignment and self-folding is that strong structural consensus among related, but diverged sequences are a good indicator for preserved functional role. Up to now there is no a general solution for this long standing problem.

Currently available models (foldalign (13), dynalign (15) and others) implement only restricted versions of Sankoff's program (variations around first align and then fold or oppositely).

 DOI: 10.5504/50YRTIMB.2011.0039

Simultaneous optimization of alignment and self-folding requires these two otherwise different in their nature processes, to be based on a common theoretical frame. Here we propose such common theoretical frame based on the analogy between the classical alignment algorithm and hybridization between two nucleotide sequences without self-folding.

Hybridization

Consider a pair of sequences A=$\{A_1, \ldots., A_i, \ldots. A_N\}$ and B=$\{B_1, \ldots., B_i, \ldots. B_M\}$, 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 RNA/DNA $\mathcal{A}$ ={A, T(U), G, C}. The hybridization between A and B is based on the condition that there are at least two nucleotides (A_i, B_j) that are in contact. Sequence enumeration is always from the $5'-$ to $3'-$ end of sequences. The contact (A_i, B_j) includes the initiation free energy term necessary to bring the two sequences together (4).

With increasing of the temperature the overwhelming majority of the double-stranded conformations of sequences A and B tend toward their corresponding unfolded states. At each temperature there is an ensemble of conformational states where each conformation is characterized with the fraction of its base pairs and their location along the double-stranded hybridization form of sequences A and B, which are melted at that given temperature. Thus along the sequences we have variety of local structural motifs characterized by alternating loops -single stranded regions- and double stranded regions. The location and the length of these local structural motifs depend on their relative Boltzmann statistical weights.

Energy rules for hybridization and self-folding are based on the assumption that stacking base pairs and loop entropies contribute additively to the free energy of a nucleic acid secondary structure (4).

Comparison of short RNAs/DNAs with different base pairs, loop sequences, bulges, etc. has yielded an extremely useful database of thermodynamic parameters from which the stabilities of conformational states of larger nucleic acid sequences can be estimated. The estimation of the thermodynamic parameters is based on the nearest-neighbor approximation for inter-residue interactions of the closest along the sequence nucleotide residues (9, 19).

Therefore, in the standard energy model secondary structure is divided into loops, and a free energy is associated with every loop. The total free energy is the sum of loop free energies. The standard model consists of the following loop types: stacking, hairpin, bulge, internal loop and multibranched loops (4). There are also duplex ends (**Fig. 1**).

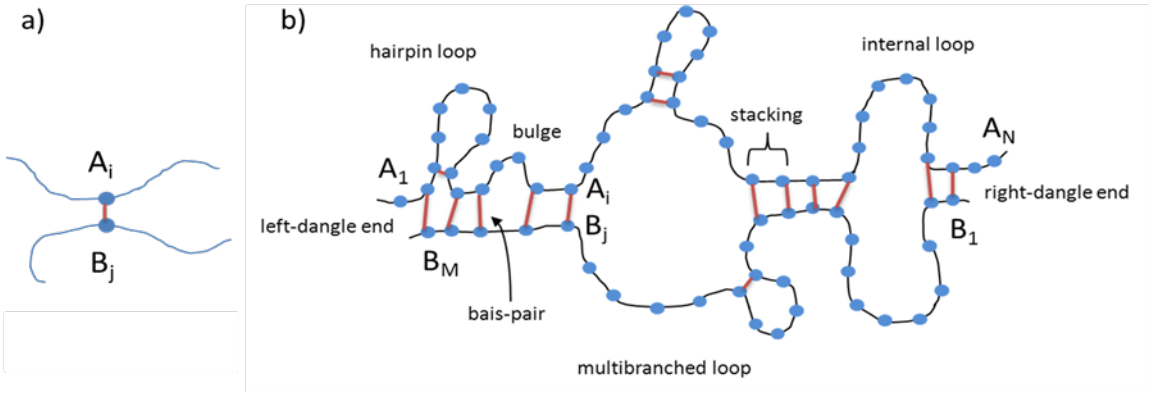

Fig. 1. a) Initiation of hybridization is based on the condition that at least there are two nucleotides (A_i, B_j) along sequences A and B that are in contact. b) Loop types in the standard model for hybridization

Alignment

Consider now that instead of hybridization we want to align the sequences A and B. An alignment represents 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 (5, 6).

Let (as shown in **Fig. 2**) $p_{A_iB_j}$ 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:

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

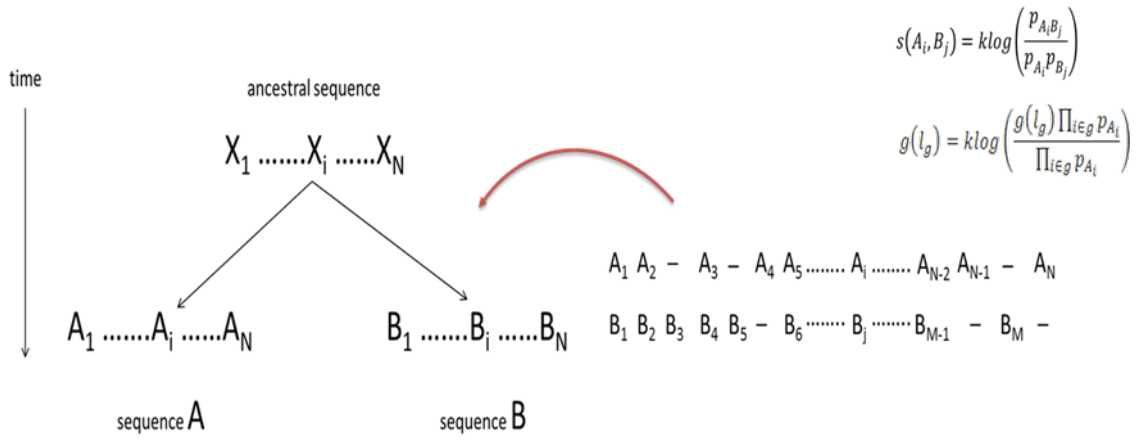

Fig. 2. Alignment represents a specific hypothesis about the evolution of the sequences

There are two main approaches to determine the parameters in the scoring model: 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.; 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 stochastic process of nucleotide substitution (6).

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 l_g and the combined probability of the set of inserted nucleotides $\prod_{i\in g} p_{A_i}$ (1, 6).

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

Loop scoring model for alignment

Simultaneous optimization of alignment and self-folding requires these two otherwise different in their nature processes, to be based on a common theoretical frame. Here we propose such common theoretical frame based on the analogy between the classical alignment algorithm and hybridization between two nucleotide sequences.

In order to find a correspondence between hybridization energy rules and alignment scoring rules we need to divide the alignment between two sequences into loops and assigned a score to each loop in such way that the total score of the alignment is a sum over the scores for each alignment loop. Based on the analogy with the hybridization we consider the following loops: stacking, bulge, internal loops and dangle ends. Hairpins and multibranched loops which take into account intramolecular basepairs are not considered (**Fig. 3**).

The score for the alignment $\mathcal{A}$ 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:

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

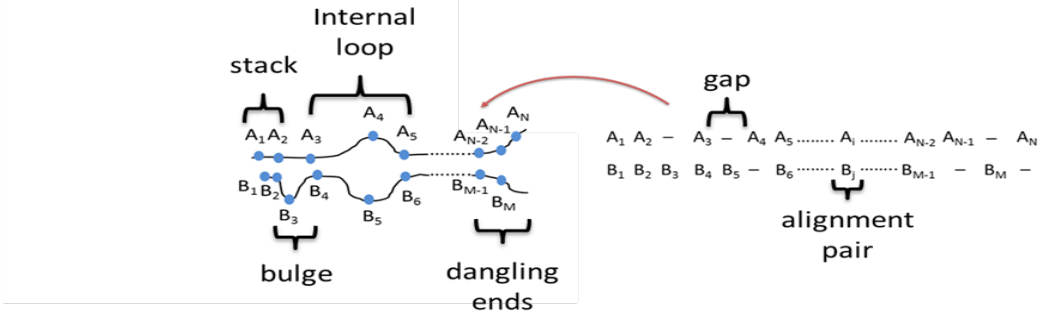

Fig. 3. Correspondents between the matched, mismatched and gap alignment scoring and loop based scoring

Fig. 3 shows how to group scoring terms into loop groups:

stacking loop- $\binom{A_1A_2}{B_1B_2} = log\left(\frac{p_{A_1B_1}}{p_{A_1}p_{B_1}}\right) + log\left(\frac{p_{A_2B_2}}{p_{A_2}p_{B_2}}\right)$;
bulge loop- $\binom{A_2\ \_\ A_3}{B_2B_3B_4} = log\left(\frac{p_{A_2B_2}}{p_{A_2}p_{B_2}}\right) + g\left(\binom{\_}{B_3}\right) + log\left(\frac{p_{A_3B_4}}{p_{A_3}p_{B_4}}\right)$;
internal loop- $\binom{A_3\ \_\ A_4A_5}{B_4B_5 - B_6} = log\left(\frac{p_{A_3B_4}}{p_{A_3}p_{B_4}}\right) + g\left(\binom{\_}{B_5}\right) + g\left(\binom{A_4}{\_}\right) + log\left(\frac{p_{A_5B_6}}{p_{A_5}p_{B_6}}\right)$
dangling end - $\binom{A_{N-2}\ A_{N-1}\ \_\ A_N}{B_{M-1}\ \_\ B_M\ \_} = log\left(\frac{p_{A_{N-2}B_{M-1}}}{p_{A_{N-2}}p_{M-1}}\right) + g\left(\binom{A_{N-1}}{\_}\right) + g\left(\binom{\_}{B_M}\right) + g\left(\binom{A_N}{\_}\right)$.

The grouping of the score terms is based on the condition that one does not count gap permutations for example such as:
$\binom{\_\ A_4}{B_5\ \_} = g\left(\binom{\_}{B_5}\right) + g\left(\binom{A_4}{\_}\right)$ and $\binom{A_4\ \_}{\_\ B_5} = g\left(\binom{A_4}{\_}\right) + g\left(\binom{\_}{B_5}\right)$

Partition function

Computation of the partition function for alignments was pioneered by Miyazawa (5, 16). For the probability of a particular alignment between the sequences A and B we can write:

$$P(\boldsymbol{\mathcal{A}}) \approx e^{\frac{S(\boldsymbol{\mathcal{A}})}{k}}$$

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

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

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

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

where the parameter T plays the role of temperature.

The partition function is computed by means of a recursion calculation algorithm we have previously developed for hybridization between two sequences. This approach is based on the additive property of the energy rules for hybridization (4).

Thus, each nucleotide pair $(A_i - B_j)$ formally divides the hybridized form AB of the sequences A and B in two parts- left L from position $(1, M)$ to position (i, j) and right R from position $(N, 1)$ to position (i, j) – in such way that the free energy $F(A_i, B_j)$ of AB is a sum of the free energies of the left $FL(A_i, B_j)$ and right $FR(A_i, B_j)$ parts plus the free energy $F^{initiation}$ of initiation of hybridization process (**Fig. 4**).

$$F(AB) = FL(A_i, B_j) + FR(A_i, B_j) + F^{initiation}$$

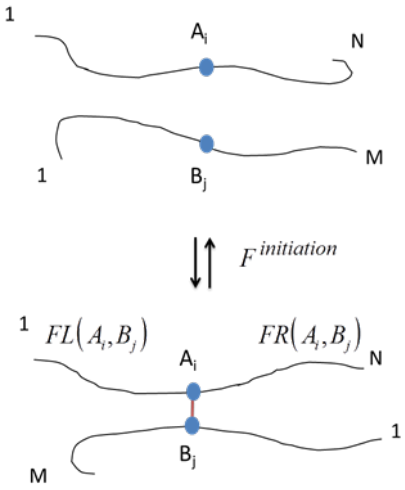

Fig. 4. Additive property of the free energy rules for hybridization based on the nearest-neighbor approximation (4)

The additivity of the free energy leads to a multiplication of the partition functions of the left $ZL(i, j)$ and right $ZR(i, j)$ parts. Thus, $ZL(i, j)$ is the partial partition function from position $(1, M)$ to position (i, j) and $ZR(i, j)$ is the partial partition function from position $(N, 1)$ to position (i, j) (**Fig. 5**).

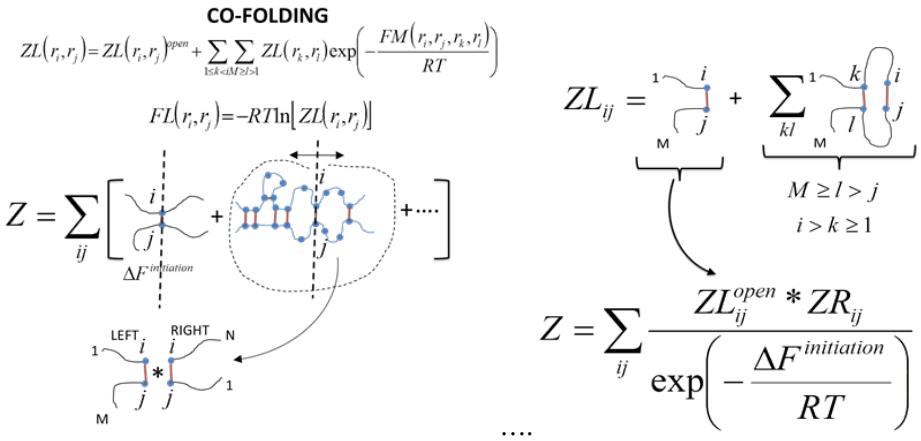

Fig. 5. The partition function is computed by means of a recursion calculations algorithm we have previously developed for hybridization between two sequences (4)

By analogy with hybridization and based on loop scoring rules for each alignment $\boldsymbol{\mathcal{A}}$ the score of the whole alignment is the sum of the score of the partial alignment $SL(i,j)$ from position (1, 1) to position (i,j) and the score of the partial alignment $SR(i,j)$ from position (N,M) to position (i,j), minus the score of the match/mismatch (i,j), $s(A_i,B_j)$.

$$S(\boldsymbol{\mathcal{A}}) = SL(i,j) + SR(i,j) - s(A_i,B_j)$$

Partial alignments $SL(i,j)$ and $SR(i,j)$ are calculated by recursive relation algorithm 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 global optimal alignment for a given model is based on maximizing the similarity score between sequences. The best score for an alignment is the value of the final cell of the matrix S (5). 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.

The partition function is computed by means of a recursion calculations algorithm and used to determine the probability of an alignment as well as the probability of each possible match or mismatch between two sequence positions (i,j). For calculation of match/mismatch probabilities, detailed recursion relations for partition functions of alignments are based on their recursion analogs for hybridization of subsequences. The partition function is used for backtracking and reconstructing a properly weighted ensemble of optimal and suboptimal alignments.

Using the additive property of alignment score we can write for the match/mismatch probability (4):

$$P(A_i,B_j) = \frac{ZL^{open}(A_i,B_j)ZR(A_i,B_j)}{Ze^{\frac{S(A_i,B_j)}{kT}}}$$

Application and detailed analysis of the model is postponed to a future article.

Conclusions

RNA/DNA molecules 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. This makes homology assessment questionable or impossible as well as possible lost of phylogenetic signals.

The objective of our work was to present a general method 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. Up to now, there was no a general solution for this long standing problem.

Simultaneous optimization of alignment and self-folding requires these two otherwise different in their nature processes, to be based on a common theoretical frame. Here we presented such common theoretical frame based on the analogy between the

classical alignment algorithm and hybridization between two nucleotide sequences without self-folding. Using this analogy we solved the problem of simultaneous prediction of secondary structure and alignment between nucleotide sequences by dividing the Sankoff's program in two steps.

In this article we presented solution to the first step, i.e. incorporated in the alignment algorithm a loop-based scoring schema with analogy to the full loop-based RNA/DNA energy model for hybridization of two sequences. In order to find a correspondence between hybridization energy rules and alignment scoring rules we divided the alignment between two sequences into loops and assigned a score to each loop in such way that the total score of alignment is a sum over the scores for each alignment loop. Based on analogy with hybridization we considered the following loops: stacking, bulge, internal loops and dangle ends.

Additive property of obtained scoring loop rules allowed us to calculate alignment partition function using our results obtained previously for hybridization between two sequences. Thus, calculation of the thermodynamic partition function over all possible double-stranded conformations is interpreted in terms of all possible canonical pairwise alignments; detailed recursion relations for partition functions of alignments are based on their recursion analogs for hybridization of subsequences.

In this work we did not consider self-folding of the aligned sequences. In our next article we will include in the alignment an algorithm for self-folding which will give the full solution of Sankoff's program.

Acknowledgements

This work was supported by grant **D002-176 TK** from National Science Foundation of the Republic of Bulgaria for which the authors are grateful.

REFERENCES

1. **Altschul S.F. and Erickson B.W.** (1986) Bull. Math. Biol., **48**(5-6), 603-616.
2. **Blake J.D. and Cohen F.E.** (2001) J. Mol. Biol., **307**(2), 721-735.
3. **Chothia C. and Lesk A.M.** (1986) EMBO J., **5**, 823-826.
4. **Dimitrov R.A.** (2005) Bulg. J. Phys., **32**, 220-235.
5. **Dimitrov R.A. and Gouliamova D.E.** (2012) Biotechnology and Biotechnology Equipment (in press).
6. **Durbin R., Eddy S., Krogh A., Mitchison G.** (1998) Biological sequence analysis, Probabilistic models of proteins and nucleic acids, Cambridge University Press.
7. **Felsenstein J.** (2004) Inferring Phylogenies, Sunderland MA, Sinauer Associates.
8. **Fitch W.M. and Ayala F.** (1994) PNAS, **91**, 6717-6720.
9. **Freier S.M., Alkema D., Sinclair A., Neilson T., Turner D.H.** (1983) Biochemistry, **22**, 6198-6206.
10. **Gardner P.P., Wilm A., Washietl S.** (2005) Nucleic Acids Research, **33**, 2433-2439.
11. **Ginalski K.** (2006) Curr. Opin. Struct. Biol., **16**(2), 172-177.
12. **Gregory T.R.** (2004) Gene, **324**, 15-34.
13. **Havgaard J.H., Lyngso R.B., Gorodkin J.,** (2005) Nucleic Acids Research, W650-653.
14. **Kaczanowski S. and Zielenkiewicz P.** (2010) Theoretical Chemistry Accounts, **125**, 543-550.
15. **Mathews D.H. and Turner D.H.** (2002) J. Mol. Biol., **317**(2), 191-203.
16. **Miyazawa S.** (1994) Protein Eng., **8**(10), 999-1009.
17. **Nee S., Mooers A.O., Harvey P.H.** (1992) PNAS, **89**, 8322-8326.
18. **Sankoff D.** (1985) SIAM Journal of Applied Mathematics, **45**, 810-825.
19. **Sugimoto N., Kierzek R., Turner D.H.,** (1987) Biochemistry, **26**, 4554-4558.
20. **Wass M.N. and Sternberg M.J.E.** Bioinformatics, (2008) **24**, 798-806.