*Genes, Genomes and Genomics* *©2009 Global Science Books*

# Assessment of Molecular (Dis)similarity: The Role of Multiple Sequence Alignment (MSA) Programs in Biological Research

**Ahmed Mansour[1*] • Jaime A Teixeira da Silva[2] • Gábor Gyulai[3]**

[1] Genetics Department, Faculty of Agriculture, Zagazig University, Zagazig, Egypt
[2] Faculty of Agriculture and Graduate School of Agriculture, Kagawa University, Miki-cho, Kagawa 761-0795, Japan
[3] Institute of Genetics and Biotechnology, St. Stephanus University, Gödöllő, H-2103, Hungary

*Corresponding author*: * **alzohairy@yahoo.com** or **amansour@zu.edu.eg**

## ABSTRACT

Multiple sequence alignments (MSAs) are powerful tools in modern molecular biology that rely on sequence comparison methods. Based on MSAs, structural models, functional predictions, and phylogenetic trees can be created. MSAs are also used to infer the function of newly sequenced genes, predict new members of gene families, explore evolutionary relationships, sequence annotation, structural and functional predictions for genes and proteins. In this review, we emphasize the practical application of different MSA methods.

**Keywords:** bioinformatics, genetic diversity

## INTRODUCTION

A multiple sequence alignment (MSA) is a sequence alignment tool of three or more sequences of protein, DNA, or RNA MSA programs can be used to conduct phylogenetic analysis to assess shared sequences and predict evolutionary lineages (Gotoh 1993). Sequence alignments, that are well characterized, can reveal structure and function of the studied sequences.

MSAs are used to explore important biological domains and motifs within a sequence by extracting consensus sequences of SSRs that may have essential roles in the structure and function of certain genes or proteins (Feng and Doolittle 1987; Taylor 1988; Madhusudhan *et al.* 2009) (**Fig. 1A**).

MSAs have been used to provide insight into the function of newly sequenced genes and predict new members of a gene family through sequence homology, structure similarity and orthology (Lipman *et al.* 1989; Garza-Garcia *et al.* 2009), and evolutionary relationships (Glazer and Kechris 2009). They were also used for sequence annotation (Galat 2009), structural and functional predictions for genes and proteins with a final conclusion in biological processes and phylogeny (Szabó *et al.* 2005; Gyulai *et al.* 2006; Kocabıyık and Demirok 2009) (**Fig. 1B, 1C**).

### The four major steps in MSA programs

MSA programs arrange sequences of amino acids or nucleotides under the same column according to their similarity and homology. In theory, homologous sequences share a common ancestor and usually also share common functions.

MSA programs include four major successive steps which include: 1) data mining of DNA, RNA or protein sequences, 2) inputting data into an automatic MSA program, 3) editing the resulting alignments and, finally 4) interpreting the results (**Fig. 2**). It is important practically, and advisable to check the running alignment by adding a few distantly related sequences, "marginal hits", one by one to examine the effect of these sequences on the overall alignment quality (Higgens and Taylor 2000).

### MSA adds a new dimension to proteomics research

In proteomics, MSA is a powerful computational tool for the molecular biologist in detecting conserved regions and motifs among distantly related proteins. For instance, it was used to detect evolutionary change in proteins (Dayhoff *et al.* 1978) and to identify compensating changes between residues at positions in protein multiple sequence alignment (Taylor and Hatrick 1994). In most cases, when sequences are similar, their structure, evolution, and function are most probably closely related. Thus, similar protein domains provide useful information for predicting protein structure, function, evolution and design since similar protein sequences may contain similar known domains (Kalsum *et al.* 2009). In this regard, survival of initial patterns after a second alignment indicates that this conserved pattern alignment may be true.

In 1994, a new method was developed by Taylor and his colleagues to compare two different protein structures and to combine them into a multiple structure consensus (Taylor *et al.* 1994). This method enhanced the quality of the resulting sequence alignments and phylogenetic tree construction. However, there are some other pattern-identifying tools that could verify the existence of these conserved patterns such as ScanProsite (Hulo *et al.* 2008) and MotifScan (Chen *et al.* 2002) which is hyperlinked in the ExPASy Proteomics Server Site (http://ca.expasy.org/). Recently, new software called SplitSSI-SVM was developed to reduce misleading false positive aligments and increase the strength of domain signal sensitivity and specificity on single-domain, two-domains and multiple-domains over other protein domain predictors (Kalsum *et al.* 2009).

### Converting a protein's MSA to the corresponding DNA

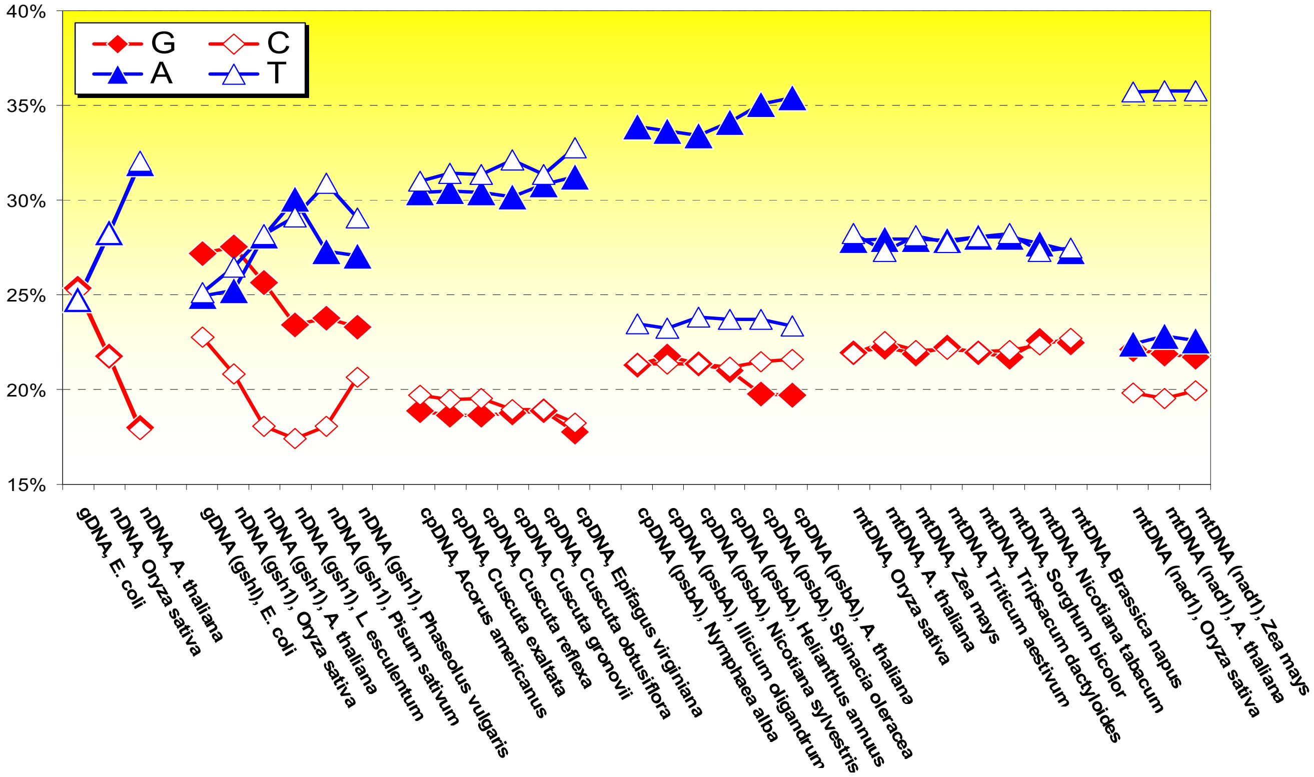
Most of the MSA programs, as described above, are only effective when sequences of similar type are compared (e.g. proteins with proteins or DNA with DNA). However, comparing protein with its coding DNA is needed sometimes (Claverie and Notredame 2007). This kind of alignment

Received: 20 May, 2009. Accepted: 26 October, 2009.

**A**

```
 cultivars/ bp                ....... ...|....|.....|....|....|....|....|....|....|....|....|....|....|....|....|..
                              45        55        65        75        85        95        105       115       125
[ctt]$_{20}$ cta[ctt]$_{4}$
 25. Kosarga                  ggggggc cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctacttcttcttcttca
 28. Hales-Best               ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctacttcttcttctt..
[ctt]$_{19}$cta[ctt]$_{4}$
 33. Hegykoi(a)               ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctt---ctacttcttcttctt..
 35. Kisteleki                ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctt---ctacttcttcttctt..
 36. Nagycserkeszi            ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctt---ctacttcttcttctt..
 37. Nyírbatori (a)           ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctt---ctacttcttcttctt..
 41. Fortuna                  ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctt---ctacttcttcttctt..
 43. Taktaharkanyi            ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctt---ctacttcttcttctt..
[ctt]$_{18}$cta[ctt]$_{4}$
 44. Sarandi (a)              ....... cttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttcttctt------ctacttcttcttctt..
[ctt]$_{14}$cta[ctt]$_{4}$
 12. Pusztadobosi             ....... cttcttcttcttcttcttcttcttcttcttcttcttcttctt------------------ctacttcttcttctt..
[ctt]$_{11}$cta[ctt]$_{4}$
 ◆ 15th CENT.                 ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
  2. Sweet-Ananász            ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
  3. Nyíribronyi              ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
  4. Javított-Zentai          ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
  6. Hevesi                   ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
  7. Ezust-Ananasz            ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 12. Pusztadobosi             ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 13. Muskotaly                ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 15. Topaz                    ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 17. Kiskorosi                ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 21. Túrkevei                 ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 36. Nagycserkeszi            ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 41. Fortuna                  ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 44. Sarandi (b)              ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
 46. Magyar kincs (a)         ....... cttcttcttcttcttcttcttcttcttcttctt---------------------------ctacttcttcttctt..
[ctt]$_{8}$cta[ctt]$_{4}$
  5.  Turai                   ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 14. Csardaszallasi           ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 19. Soponyai                 ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 20. Penyigei                 ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 24. Hogolyo                  ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 26. Tarnamerai               ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 30. Tetenyi-Csh.             ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 31. Muhi                     ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 33. Hegykoi (b)              ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 35. Kisteleki                ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 37. Nyírbatori (b)           ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 42. Kallosemjeni             ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
 46. Magyar kincs (b)         ....... cttcttcttcttcttcttcttctt------------------------------------ctacttcttcttctt..
[ctt]$_{7}$cta[ctt]$_{4}$
 38. Szirmai                  ....... cttcttcttcttcttcttctt---------------------------------------ctacttcttcttctt..
```

**B**



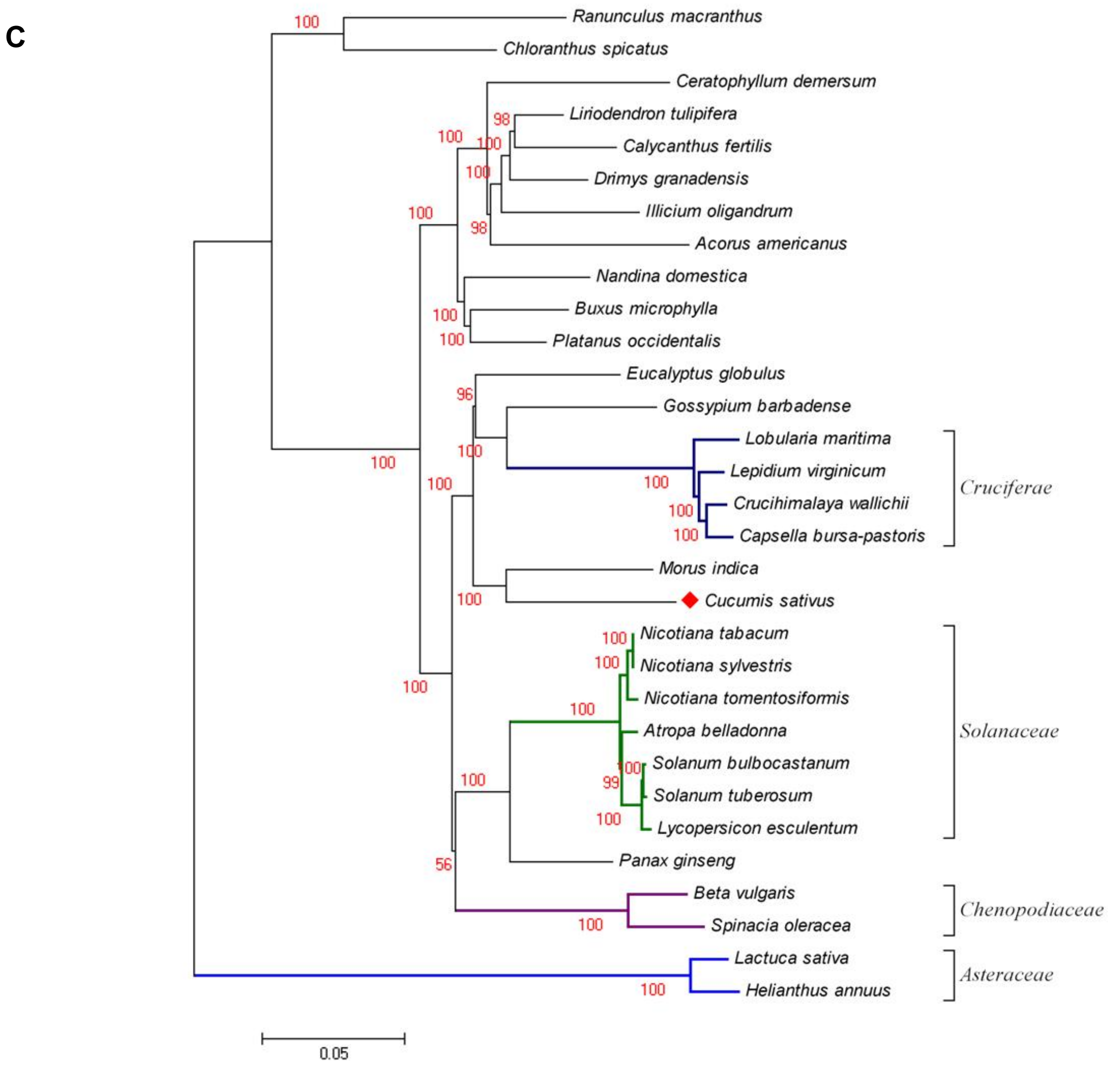


**Fig. 1 Three MSA samples.** (**A**) Samples of MSA analysis for the compound microsatellite SSR alleles at the (CTT)n-CTA-(CTT)4 locus (45–15 bp) of ancient (15th century) (Szabó *et al*. 2005) and modern melon cultivars (*Cucumis melo*). Arrows indicate the shortening length of SSRs (CTT)$_{20}$ → (CTT)$_{8.}$ Changes in SSR alleles length were associated with the evolutionary time of each cultivar. (**B**) Samples of comparative nucleotide (G, C, A, and T) analysis (%) of the DNA of total genomic (gDNA), nuclear (nDNA) and organelle (cpDNA and mtDNA) samples compared to the sequences of coding DNAs (*gsh*I, *gsh*1, *psb*A and *nad*1). Data show shift to high A-content in the coding region of cpDNAs and high T-content in coding mtDNAs. See also the shift from *E. coli* to *A. thaliana*. Accession numbers: NC_003070, NC_008402.1, NC_003076.5, X03954, AJ508916, Y09944, AF017983, AF128455, AF128454, NC_010093.1, NC_009963.1, NC_009766.1, NC_009765.1, NC_009949.1, NC_001568.1, NC_006050, EF380354, AB237912, DQ383815, AJ400848, AP000423, NC_007886.1, NC001284.2, NC_007982.1, NC_007579.1, NC_008362.1, NC_008360.1, NC_006581.1, NC_008285, NC_007886, NC_001284, NC_007982. (**C**) Samples of the application of MSA including BLAST computing by MULTALINE server combined with dendrogram analysis (with bootstrap values) using MEGA4 program (Tamura *et al.* 2007). Total DNA sequences of cpDNAs (4.678.893 nt in total) of 31 species were analyzed. Accession numbers: *Acorus americanus* (EU273602, 153.819 nt, Acoraceae), *Atropa belladonna* (AJ316582, 156.687 nt, Solanaceae), *Beta vulgaris* (EF534108, 149.696 nt, Chenopodiaceae), *Buxus microphylla* (EF380351, 159.010 nt, Buxaceae), *Calycanthus fertilis* (AJ42841, 153.337 nt, Calycanthacea*e*), *Capsella bursa-pastoris* (AP009371, 154490 nt, Cruciferae), *Ceratophyllum demersum* (EF614270, 156252 nt, Ceratophyllaceae), *Chloranthus spicatus* (EF380352, 157.772 nt, Chloranthaceae), *Crucihimalaya wallichii* (AP009372, 155.199 nt, Cruciferae), *Cucumis sativus* (AJ970307, 155.293 nt, Cucurbitaceae), *Drimys granadensis* (DQ887676, 160.604 nt, Winteraceae), *Eucalyptus globulus* (AY780259, 160.286 nt, Myrtaceae), *Gossypium barbadense* (AP00912, 160.317 nt, Malvaceae), *Helianthus annuus* (DQ383815, 151.104 nt, Asteraceae), *Illicium oligandrum* (EF380354, 148.553 nt, Schisandraceae), *Lactuca sativa* (DQ383816, 152772 nt, Asteraceae), *Lepidium virginicum* (AP009374, 154.743 nt, Cruciferae), *Liriodendron tulipifera* (DQ899947, 159.886 nt, Magnoliaceae), *Lobularia maritima* (AP009375, 152.659 nt, Cruciferae), *Lycopersicon esculentum* (DQ347959, 155.461 nt, Solanaceae), *Morus indica* (DQ226511. 158.484 nt, Moraceae), *Nandina domestica* (DQ923117, 156.599 nt, Berberidaceae), *Nicotiana sylvestris* (AB237912, 155.941 nt, Solanaceae), *Nicotiana tabacum* (Z00044, 155943 nt, Solanaceae), *Nicotiana tomentosiformis* (AB240139, 155.745 nt, Solanaceae), *Panax ginseng* (AY582139, 149.696 nt, Araliaceae), *Platanus occidentalis* (DQ923116, 161.791 nt, Platanaceae), *Ranunculus macranthus* (DQ359689, 155.129 nt, Ranunculaceae), *Solanum bulbocastanum* (DQ347958, 155.371 nt, Solanaceae), *Solanum tuberosum* (DQ386163, 155.298 nt, Solanaceae), *Spinacia oleracea* (AJ400848, 150.725 nt, Chenopodiaceae). **Fig. 1A-C** unpublished.

requires insertion of long gaps corresponding to the introns with special tools and software. This can be very useful to deal with frame shifts in the input alignment, which is usually the case in the analysis of pseudogenes. Such codon alignments can be also used to evaluate the type and rate of nucleotide substitutions in coding DNA for a wide range of evolutionary analyses, such as the identification of levels of selective constraint acting on genes, or to perform DNA-based phylogenetic studies. *PAL2NAL* is one of the rare programs that can carry out this task (http://www.bork.embl.de/pal2nal/) by converting an MSA of proteins and the corresponding DNA (or mRNA) sequences into a codon alignment (Suyama *et al.* 2006). This Program automatically assigns the corresponding codon sequence, even if the input DNA sequence has mismatches with the input protein sequence, or contains UTRs, polyA tails. A new mirror site for PAL2NAL was established at Kyoto University at (http://www.genome.med.kyoto-u.ac.jp/cgi-bin/suyama/pal2nal/

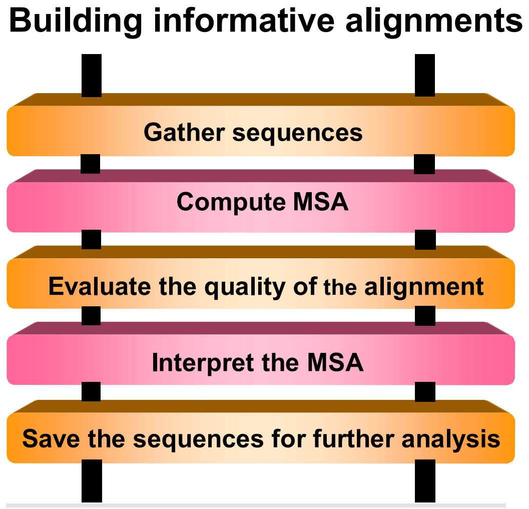

Building informative alignments
Gather sequences
Compute MSA
Evaluate the quality of the alignment
Interpret the MSA
Save the sequences for further analysis

**Fig. 2 The general steps of MSA.** Analysis with MSA programs includes four basic steps: 1) data mining of DNA, RNA or protein sequences, 2) input into an automatic MSA program, 3) editing the alignments, and finally 4) interpreting the results.

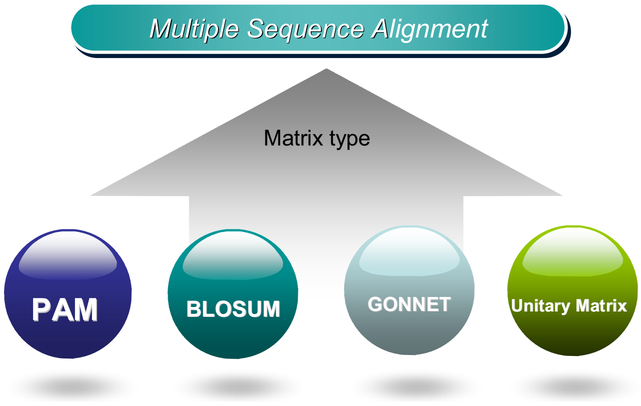


**Fig. 3 Four types of matrices are used for comparative analyses.** PAM (Point or Percent Accepted Mutation, BLOSUM (Blocks Substitution Matrix), GONNET matrix, and the DNA Identity Matrix (Unitary Matrix).

index.cgi).

*Protal2dna* is another program which aligns DNA sequences corresponding to a protein alignment (http://bioweb2.pasteur.fr/soft-pasteur.html#protal2dna). Both servers require protein and the corresponding DNA before use. However, if a protein sequence is only available, running a Blastx against a complete genome or using a Web server called Protogene (www.tcoffee.org) fetches the DNA sequence corresponding to a given protein (Moretti *et al.* 2006).

## Understanding the theories underlying a given MSA matrix

The choice of a scoring function that reflects biological or statistical observations about known sequences is important to producing good alignments. There are mainly four types of matrices used for comparative analyses, which is PAM (Point or Percent Accepted Mutation), BLOSUM (Blocks Substitution Matrix), GONNET matrix and the DNA Identity Matrix (Unitary Matrix). There have been extensive studies looking at the frequencies in which amino acids substituted for each other during evolution. Many of the MSA softwares can give the choice to switch from one matrix to another like ClustalW (http://www.ebi.ac.uk/Tools/clustalw2/index.html). Alignment algorithms and software can be directly compared to one another using a standardized set of benchmark reference multiple sequence alignments known as BAliBASE (http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/prog_scores.html) (Thompson *et al.* 1999). Understanding a given matrix makes it possible to choose the most useful program for the designed study (**Fig. 3**). In simple terms, a substitution matrix describes the likelihood that two residue types would mutate to each other in evolutionary time. Theories underlying a given matrix will be explained in more detail in the next sections.

## PAM matrices

Methods for alignment of protein sequences typically measure similarity by using a substitution matrix with scores for all possible exchanges of one amino acid with another. Based on 1572 observed mutations in 71 families of closely related proteins, Margaret Dayhoff has introduced PAM matrices in 1978. Point accepted mutation (PAM) was developed primarily for scoring amino acid matrices, which refer to various degrees of sensitivity depending on the evolutionary distance among the sequences studied (Dayhoff *et al.* 1978). PAM matrices are a set of matrices used to score sequence alignments and each matrix is twenty-by-twenty (for the twenty standard amino acids); the value in a given cell represents the probability of a substitution of one amino acid for another (**Table 1**). Simply, PAM40, for example, means 40 mutations per 100 amino acids in the sequence. Many amino acid replacement matrices were proposed to improve the general matrix (Le and Gascuel 2008). The PAM matrices imply a Markov chain model of protein mutation (Baldi *et al.* 1994). Scores from a PAM matrix were deployed for the identification of amino acid substitutions while detecting distant repeats. Analysis of the distant repeats would enable the establishment of a firm relationship with the repeats with respect to their function and three-dimensional structure during the evolutionary process (Banerjee *et al.* 2008).

## BLOSUM matrices

BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) (Henikoff and Henikoff 1992) is a substitution matrix used for sequence alignments of proteins (the default in blastp). Several sets of BLOSUM exist using different alignment databases, named with numbers (**Table 2**). The percentage used was appended to the name, giving BLOSUM45, for instance, where sequences that were more than 45% identical were clustered. BLOSUM with high numbers (e.g. BLOSUM80) are designed for comparing closely related sequences at the 80% identity level which can be used for less divergent alignments. However, BLOSUM with low numbers (e.g. BLOSUM45) are consequently designed for comparing distantly related sequences at the 45% identity level which can be used for more divergent alignments. BLOSUM matrices are most sensitive for local alignment of related sequences and therefore are ideal for identifying an unknown sequence. Recently, it was shown that calculation in BLOSUM62 was not exactly accurate; however, this miscalculation improved the search performance and results (Styczynski *et al.* 2008). This matrix is included in many MSA programs such as ClustalW (http://www.clustal.org/).

## GONNET matrix

This matrix uses an exhaustive pairwise alignment to measure differences among amino acid composition of proteins (Gonnet *et al.* 1992). This model provides theoretical support for using insertions and deletions (indels) as part of "parsing algorithms", important in the *de novo* prediction of the folded structure of proteins from sequence data (Benner *et al.* 1993). This weighted matrix, based on the supposition that any given amino acid substitution is influenced by neighboring amino acids, provides a basis for characterizing protein families (**Table 3**). This matrix has been used in many research projects recently, for instance to study clus-

**Table 1** The PAM 250 matrix. This is appropriate for searching for alignments of sequence that have diverged by 250 PAMs, 250 mutations per 100 amino acids of sequence. Because of back mutations and silent mutations this corresponds to sequences that are about 20 percent identical (http://www.ebi.ac.uk/2can/tutorials/matrices)

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 12 | | | | | | | | | | | | | | | | | | | |
| G | -3 | 5 | | | | | | | | | | | | | | | | | | |
| P | -3 | -1 | 6 | | | | | | | | | | | | | | | | | |
| S | 0 | 1 | 1 | 1 | | | | | | | | | | | | | | | | |
| A | -2 | 1 | 1 | 1 | 2 | | | | | | | | | | | | | | | |
| T | -2 | 0 | 0 | 1 | 1 | 3 | | | | | | | | | | | | | | |
| D | -5 | 1 | -1 | 0 | 0 | 0 | 4 | | | | | | | | | | | | | |
| E | -5 | 0 | -1 | 0 | 0 | 0 | 3 | 4 | | | | | | | | | | | | |
| N | -4 | 0 | -1 | 1 | 0 | 0 | 2 | 1 | 2 | | | | | | | | | | | |
| Q | -5 | -1 | 0 | -1 | 0 | -1 | 2 | 2 | 1 | 4 | | | | | | | | | | |
| H | -3 | -2 | 0 | -1 | -1 | -1 | 1 | 1 | 2 | 3 | 6 | | | | | | | | | |
| K | -5 | -2 | -1 | 0 | -1 | 0 | 0 | 0 | 1 | 1 | 5 | 5 | | | | | | | | |
| R | -4 | -3 | 0 | 0 | -2 | -1 | -1 | -1 | 0 | 1 | 2 | 3 | 6 | | | | | | | |
| V | -2 | -1 | -1 | -1 | 0 | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 4 | | | | | | |
| M | -5 | -3 | -2 | -2 | -1 | -1 | -3 | -2 | 0 | -1 | -2 | 0 | 0 | 2 | 6 | | | | | |
| I | -2 | -3 | -2 | -1 | -1 | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 4 | 2 | 5 | | | | |
| L | -6 | -4 | -3 | -3 | -2 | -2 | -4 | -3 | -3 | -2 | -2 | -3 | -3 | 2 | 4 | 2 | 6 | | | |
| F | -4 | -5 | -5 | -3 | -4 | -3 | -6 | -5 | -4 | -5 | -2 | -5 | -4 | -1 | 0 | 1 | 2 | 9 | | |
| Y | 0 | -5 | -5 | -3 | -3 | -3 | -4 | -4 | -2 | -4 | 0 | -4 | -5 | -2 | -2 | -1 | -1 | 7 | 10 | |
| W | -8 | -7 | -6 | -2 | -6 | -5 | -7 | -7 | -4 | -5 | -3 | -3 | 2 | -6 | -4 | -5 | -2 | 0 | 0 | 17 |

**Table 2** Blosum 45 Matrix. This is derived from sequence blocks clustered at the 45% identity level.

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 7 | | | | | | | | | | | | | | | | | | | |
| P | -2 | 9 | | | | | | | | | | | | | | | | | | |
| D | -1 | -1 | 7 | | | | | | | | | | | | | | | | | |
| E | -2 | 0 | 2 | 6 | | | | | | | | | | | | | | | | |
| N | 0 | -2 | 2 | 0 | 6 | | | | | | | | | | | | | | | |
| H | -2 | -2 | 0 | 0 | 1 | 10 | | | | | | | | | | | | | | |
| Q | -2 | -1 | 0 | 2 | 0 | 1 | 6 | | | | | | | | | | | | | |
| K | -2 | -1 | 0 | 1 | 0 | -1 | 1 | 5 | | | | | | | | | | | | |
| R | -2 | -2 | -1 | 0 | 0 | 0 | 0 | 3 | 7 | | | | | | | | | | | |
| S | 0 | -1 | 0 | 0 | 1 | -1 | 0 | -1 | -1 | 4 | | | | | | | | | | |
| T | -2 | -1 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | 2 | 5 | | | | | | | | | |
| A | 0 | -1 | -2 | -1 | -1 | -2 | -1 | -1 | -2 | 1 | 0 | 5 | | | | | | | | |
| M | -2 | 2 | -3 | -2 | -2 | 0 | 0 | -1 | -1 | -2 | -1 | -1 | 6 | | | | | | | |
| V | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -2 | -2 | -1 | 0 | 0 | 1 | 5 | | | | | | |
| I | -4 | -2 | -4 | -3 | -2 | -3 | -2 | -3 | -3 | -2 | -1 | -1 | 2 | 3 | 5 | | | | | |
| L | -3 | -3 | -3 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | -1 | 2 | 1 | 2 | 5 | | | | |
| F | -3 | -3 | -4 | -3 | -2 | -2 | -4 | -3 | -2 | -2 | -1 | -2 | 0 | 0 | 0 | 1 | 8 | | | |
| Y | -3 | -3 | -2 | -2 | -2 | 2 | -1 | -1 | -1 | -2 | -1 | -2 | 0 | -1 | 0 | 0 | 3 | 8 | | |
| W | -2 | -3 | -4 | -3 | -4 | -3 | -2 | -2 | -2 | -4 | -3 | -2 | -2 | -3 | -2 | -2 | 1 | 3 | 15 | |
| C | -3 | -4 | -3 | -3 | -2 | -3 | -3 | -3 | -3 | -1 | -1 | -1 | -2 | -1 | -3 | -2 | -2 | -3 | -5 | 12 |
| | G | P | D | E | N | H | Q | K | R | S | T | A | M | V | I | L | F | Y | W | C |

**Table 3** An example of GONNET Matrix.

| A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | T | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.6 | 0.125 | -0.075 | 0 | -0.575 | 0.125 | -0.2 | -0.2 | -0.1 | -0.3 | -0.175 | -0.075 | 0.075 | -0.05 | -0.15 | 0.275 | 0.15 | 0.025 | -0.9 | -0.55 | A |
| | 2.875 | -0.8 | -0.75 | -0.2 | -0.5 | -0.325 | -0.275 | -0.7 | -0.375 | -0.225 | -0.45 | -0.775 | -0.6 | -0.55 | 0.025 | -0.125 | 0 | -0.25 | -0.125 | C |
| | | -0.125 | 0.675 | -1.125 | 0.025 | 0.1 | -0.95 | 0.125 | -1 | -0.75 | 0.55 | -0.175 | 0.225 | -0.075 | 0.125 | 0.125 | -0.725 | -1.3 | -0.7 | D |
| | | | 0.9 | -0.975 | -0.2 | 0.1 | -0.675 | 0.3 | -0.7 | -0.5 | -0.5 | -0.125 | 0.425 | 0.1 | 0.05 | -0.025 | -0.475 | -1.075 | -0.675 | E |
| | | | | 1.75 | -1.3 | -0.025 | 0.25 | -0.825 | 0.5 | 0.4 | -0.775 | -0.95 | -0.65 | -0.8 | -0.7 | -0.55 | 0.025 | 0.9 | 1.275 | F |
| | | | | | 1.65 | -0.35 | -1.125 | -0.275 | -1.1 | -0.875 | 0.1 | 0.1 | -0.25 | -0.25 | 0.1 | -0.275 | -0.825 | -1 | -1 | G |
| | | | | | | 1.5 | -0.55 | 0.15 | -0.475 | -0.325 | 0.3 | -0.275 | 0.3 | 0.15 | -0.05 | -0.075 | -0.5 | -0.2 | 0.55 | H |
| | | | | | | | 1 | -0.525 | 0.7 | 0.625 | -0.7 | -0.65 | -0.475 | -0.6 | -0.45 | -0.15 | 0.775 | -0.45 | -0.175 | I |
| | | | | | | | | 0.8 | -0.525 | -0.35 | 0.2 | -0.15 | 0.375 | 0.675 | 0.025 | 0.025 | -0.425 | -0.875 | -0.525 | K |
| | | | | | | | | | 1 | 0.7 | -0.75 | -0.575 | -0.4 | -0.55 | -0.525 | -0.325 | 0.45 | -0.175 | 0 | L |
| | | | | | | | | | | 1.075 | -0.55 | -0.6 | -0.25 | -0.425 | -0.35 | -0.15 | 0.4 | -0.25 | -0.05 | M |
| | | | | | | | | | | | 0.95 | -0.225 | 0.175 | 0.075 | 0.225 | 0.125 | -0.55 | -0.9 | -0.35 | N |
| | | | | | | | | | | | | 1.9 | -0.05 | -0.225 | 0.1 | 0.025 | -0.45 | -1.25 | -0.775 | P |
| | | | | | | | | | | | | | 0.675 | 0.375 | 0.05 | 0 | -0.375 | -0.675 | -0.425 | Q |
| | | | | | | | | | | | | | | 1.175 | -0.05 | -0.05 | -0.5 | -0.4 | -0.45 | R |
| | | | | | | | | | | | | | | | 0.55 | 0.375 | -0.25 | -0.825 | -0.475 | S |
| | | | | | | | | | | | | | | | | 0.625 | 0 | -0.875 | -0.475 | T |
| | | | | | | | | | | | | | | | | | 0.85 | -0.65 | -0.275 | V |
| | | | | | | | | | | | | | | | | | | 3.55 | 1.025 | W |
| | | | | | | | | | | | | | | | | | | | 1.95 | Y |

tering patterns of conserved residues in DNA-binding proteins (Ahmad *et al*. 2008). It was also used to conduct comparative genomics studies of small RNA regulatory pathway components inmosquitoes using full-length PIWI subfamily proteins and the he C-terminal 250 amino acids of Argonaute subfamily of proteins (Campbell *et al.* 2008).

**Table 4** DNA Identity Matrix (Unitary Matrix). In this matrix a positive score for a match, and a score of -10000 for a mismatch.

| | A | T | G | C |
|---|---|---|---|---|
| A | 1 | | | |
| T | -10000 | 1 | | |
| G | -10000 | -10000 | 1 | |
| C | -10000 | -10000 | -10000 | 1 |

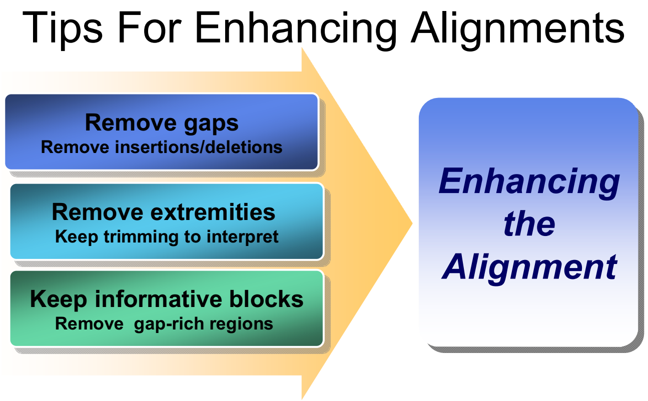


**Fig. 4 Tips for enhancing alignments.** After generating an MSA it is important to edit it manually to enhance the output.

## DNA identity matrix

In this matrix, a positive score for a match and a score of −10,000 for a mismatch is given. Since such a high penalty is given for a mismatch, no substitution should be allowed, although a gap may be permitted (**Table 4**). A penalty is subtracted for each gap introduced into an alignment because the gap increases uncertainty into an alignment (Campanella *et al.* 2003).

## Enhancing the alignments

Producing accurate and suitable multiple alignments requires that certain steps be followed (**Fig. 4**). Gaps have to be removed to get as many gap-free columns as possible. Removing the extremities beside the gap-rich regions of the alignment enhance the overall alignments of the sequences while keeping the most informative blocks (Claverie and Notredame 2007).

## Computing MSAs

There are online servers with different software that can be used for MSA analysis such as:

- ClustalW (www.ebi.ac.uk/clustalw),
- MUSCLE (phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py), Tcoffee (www.tcoffee.org),
- 3D-Coffee (http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi),
- PROBCONS (http://probcons.stanford.edu/), MAFFT (http://timpani.genome.ad.jp/Art), and
- FAST-PCR (Kalendar 2009).

## Evaluating MSAs

*Tcoffee* server is used to evaluate MSAs and remove columns that are unlikely to be correctly aligned. The most common method for *Tree* evaluation is bootstrap analysis, which essentially involves the re-sampling of the database and then an analysis of the re-sampled data. Bootstrap analysis eliminates variations introduced by re-samplings. There are some free online softwares or included into different packages that can be used for bootstrapping. For instance, *Winboot*, which is a free software, is used to estimate the statistical stability of the clusters using bootstrap analysis with 1000 replications (Yap and Nelson 1996) (http://www.irri.org/science/software/winboot.asp). Bootstrapping modules are included in many bioinformatics packages such as PHYLIP (Felsenstein 1992) or MEGA4 (Tamura *et al*. 2007; Kumar *et al*. 2008).

## Editing alignments for publishing

After generating an MSA, it is important to edit it manually to enhance the output by specifying the number of residues per line, font sizes, font styles, and colors for different regions. In this regard, there are several MSA editing programs such as Multiple Align Show (http://bioinformatics.org/sms/index.html), MultAline (http://bioinfo.genotoul.fr/multalin/), and Jalview (http://www.jalview.org/). These program text editors are specific for multiple sequence alignment. They can be either Java applet used online such as Jalview 2 (Waterhouse *et al.* 2009) or installed on the computer such as SeaView (Galtier *et al.* 1996). Jalview software can be used for editing, analysis and annotation of MSAs.

## Software sequence formats for alignment

Protein or DNA sequences need to be converted to an appropriate format for computing by MSA programs. There are many sequence formats available, such as Pearson (FASTA), Align/ClustalW (ALN/ClustalW), National Biomedical Research Foundation/Protein Information Resource (NBRF/PIR), European Molecular Biology Laboratory/Universal Protein Resource (EMBL/UniProt), Genetic Data Environment (GDE), Genetics Computer Group/Multiple Sequence Format (GCG/MSF), and Rich Sequence Format (RSF). However, most of these programs prefer the FASTA format, but "the easy way" is to use Microsoft Word to edit blocks efficiently. Recently, FASTA is the most accepted by most bioinformatics software because of its simplicity. Simply, formatting starts with the greater than symbol (>), followed by the gene/protein accession number and the sequence. However, there are many free bioinformatics softwares that can covert different formats to each other such as (SeqVerter) from (http://www.genestudio.com/seqverter.htm) or (Visual Sequence Editor) and (Readseq) from (http://iubio.bio.indiana.edu/soft/molbio/readseq/java/). There are many more online programs that can carry out the same task.

## Recognizing symbols for good and bad parts in MSA

There are some useful symbols used in most MSA programs (e.g., ClustalW, MUSCLE, Tcoffee) to tell whether a block of aligned sequences is good or not. For instance, a star (*) indicates an entirely conserved column in MSA, while a colon (:) indicates columns where all the residues have roughly the same size and the same hydropathy. However, a period (.) indicates columns where the size or the hydropathy has been preserved in the course of evolution (Claverie and Notredame 2007). For a molecular biologist, those indications are quite useful. The average *good block* is a unit at least 10–30 amino acids long, exhibiting at least one to three stars (*), a few more colons (:) close to the stars, and a several periods (.) scattered along the MSA result (**Fig. 5**). Less than 10% identity is needed in aligned multiple sequences for discovering a genuine signal. However, at least 25% identity is required to consider a pairwise alignment interesting (Ye 2008).

## New features and tools of MSA

Scalability, speed and accuracy of the MSA programs have been dramatically improved during the last few years. The new programs use new progressive alignment algorithms that are extremely fast and accurate at a modest computational cost for computing, evaluating and combining MSAs (e.g. TCOFFEE. in TCoffee package) (Poirot *et al.* 2003).

```
gb|ABW97972.1|      MGSPPPFLSKLFALVNDSYWNELIRWENNGQTFIITDPIEFSKKILPSYFKHKNFSS--F 58
gb|EDX15998.1|      ---MTPASSDLYNVN---FISEDMPTDIFEDALLPDGVEEAAKLDQQQKFGQSTVSSGKF 54
                        .*  *.*: :    : .* :  :   ::::  .  * :*    . * :...**  *

gb|ABW97972.1|      LRQLNKYGFSKLSPDEWIFGHKEFKYGKQDQLSGIIRKKKLKTNYLNFSGENIQQLNKKI 118
gb|EDX15998.1|      ASNFDVPTNSTLLDANQASTSKAAAKAQASEEEGMAVAKYSGAENGNNRDPNSSQLLR-- 112
                      :::    *.*   :     *    .: .: .*:   *   ::  *  . * .** :

gb|ABW97972.1|      EADIDFLKRSRQSFSKNFIDIYSRQEQFLIQQQNIEINQKKLESEVKILENEVCQLKGFI 178
gb|EDX15998.1|      MASVDELHGHLESMQDELETLKD-----LLRGDGVAIDQN-------MLMGALSRTPNFQ 160
                     *.:* *:   :*:..::  : .     *:: :.: *:*:       :* . :.:  .*

gb|ABW97972.1|      FGYLSKVIGKGDEKSKISDLPFFLPNEQRSSESLSKTTKNPFNFYEKFPKEEKIVFFN   236
gb|EDX15998.1|      LPEDELLLKVSTKRP---DLAFMSENSPAGGS-------NTMCFICILYQGS------   202
                    :   . ::  . ::.   **.*:  *.  ...       *.: *   : : .
```

**Fig. 5 A sequence alignment, produced by ClustalW using two sequences identified on the left by GenBank accession number.** "*": identical. ":": conserved substitutions (same colour group). ".": semi-conserved substitution (similar shapes).

In addition, some programs of protein sequence alignments incorporate three-dimensional structural information, which facilitates the alignment of distantly related sequences e.g. EXPRESSO (3DCoffee) in TCoffee package. This new feature is based on the fact that structural elements are generally more conserved than primary sequences (Armougom *et al*. 2006). For instance, comparing the structures of proteins is crucial to gaining insight into protein evolution and function (Madhusudhan *et al*. 2009). Moreover, the final alignment of MSA can now be subjected to an iterative refinement protocol and combining alternative alignment methods (e.g. MCOFFEE. in TCoffee package) (Moretti *et al*. 2007).

## Applications of MSAs

A good multiple alignment can help you discover some features for your query such as Extrapolation (to know the protein family of your sequence) (Mindnich and Penning 2009), Phylogenetic analysis (reconstruct the history of these sequences) (Shaik *et al.* 2009), Pattern identification (conserved positions characteristic of a function) (Hahn 2009), Domain identification (describes a profile for protein family PSSM) (Glazer and Kechris 2009), DNA regulatory elements (finding potentially similar binding sites), Structure prediction (prediction of protein or RNA secondary structure and in the building a 3-D model) (Bradley and Holmes 2009), nsSNP analysis (Non-Synonymous Single-Nucleotide Polymorphism) (Szabó *et al.* 2005), PCR analysis (less degenerated portions of a protein family to fish out new members by PCR) (**Fig. 6**) (Cuff and Barton 2000; Gotoh 1992; Chenna *et al*. 2003; Rausch *et al*. 2008; Waterhouse *et al*. 2009).

## Research situations where MSAs are useless

There are some research problems where multiple sequence alignments does not help to solve it. Those problems situations such as, assembling the sequence pieces in a sequencing project, turning an EST cluster into a gene sequence or where the target sequence has no homologue in any of the sequence databases (Claverie and Notredame 2007). Such problems required the use of specialized sequence assembly tools like cap3 (http://pbil.univ-lyon1.fr/cap3.php) (Huang and Madan 1999).

Picking the right sequences for MSA is very important and researchers need to know what they want to present with sequence alignment, keeping in mid the following hints: never use white spaces, never use special symbols, never use names longer than 15 characters, never give the same name to two different sequences. For instance, MSA could help in identifying important residue positions (e.g. finding conserved amino acids that are not allowed to mutate) when aligning distantly related proteins to create an extended protein family. Future improvements of MSA programs are likely to combine information of new dimensions in genetics (Pellionisz 2008).

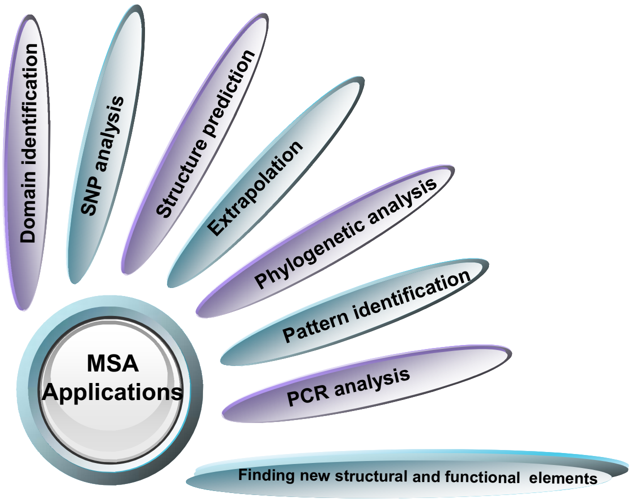


**Fig. 6 General application of MSAs.** MSAs can be used to produce structural models, functional predictions, phylogenetic trees, inferring the function of newly sequenced genes, prediction of new members of gene families, exploring evolutionary relationships, sequence annotation, structural and functional predictions for genes and proteins.

## CONCLUSIONS

MSAs were effective in predicting the location and function of genes and their transcriptional regulatory elements (Cuff and Barton 2000; Hahn 2009). MSA alignments provided structural similarity information in discovering new *Citrullus* haplotypes (Tóth *et al.* 2009) and to provide homology revealing the evolutionary triplet models of structured RNA (Bradley and Holmes 2009). MSA was used to infer functional similarity by revealing phylogeny of ALAD and MGP genes related to lead toxicity (Shaik *et al.* 2009). These programs are useful in tracking domestication and microevolution (Szabó *et al.* 2005; Gyulai *et al*. 2006), and also predicting genes that belong to a certain family, such as in revealing gene members from the aldo-keto reductase (AKR) superfamily (Mindnich and Penning 2009).

## ACKNOWLEDGEMENT

The authors thank L Waters Jr. (AU, Alabama, USA), for comments on the manuscript.

## REFERENCES

**Ahmad S, Keskin O, Sarai A, Nussinov R** (2008) Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Research* **36**, 5922-5932

**Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C** (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Research* **1 (34)**, W604-8

**Baldi P, Chauvin Y, Hunkapiller T, McClure MA** (1994) Hidden Markov models of biological primary sequence information. *Proceeding of the Na-*

*tional Academy of Sciences USA* **91**, 1059-1063

**Banerjee N, Sarani R, Ranjani CV, Sowmiya G, Michael D, Balakrishnan N, Sekar K** (2008) Algorithm to find distant repeats in a single protein sequence. *Bioinformation* **3 (1)**, 28-32

**Bradley RK, Holmes I** (2009) Evolutionary triplet models of structured RNA. *PLoS Computational Biology* **5 (8)**, e1000483

**Benner SA, Cohen MA, Gonnet GH** (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of Molecular Biology* **229**, 1065-1082

**Campanella JJ, Bitincka L, Smalley J** (2003) MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics* **10**, 4:29

**Campbell CL, Black WC 4th, Hess AM, Foy BD** (2008) Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics* **18**, 9:425

**Chen JN, Zhao X, Min JXJ, Zhang JX** (2002) Extracting biologically relevant common motifs from protein sequences. *Online course*, *Common Motifs from Multiple Proteins*. pp 41-52. Available online: http://www.cs.ubc.ca/labs/beta/Div/CPSC536A-WS-02/jchen.pdf

**Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD** (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research* **31 (13)**, 3497-500

**Claverie J, Notredame C** (2007) *Bioinformatics for Dummies* (2nd Edn), Wiley Publishing, Inc., New York, 436 pp

**Cuff JA, Barton GJ** (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics* **40 (3)**, 502-511

**Dayhoff MO, Schwartz R, Orcutt BC** (1978) A model of evolutionary change in proteins. In: *Atlas of Protein Sequence and Structure* (Vol V, 3rd Edn), National Biomedical Research Foundation, pp 345-358

**Felsenstein J** (1992) Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genetics Research* **59**, 139-147

**Feng DF, Doolittle RF** (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic tree. *Journal of Molecular Evolution* **25**, 351-360

**Galat A** (2009) On transversal hydrophobicity of some proteins and their modules. *Journal of Chemical Information and Modeling* **49**, 1821-1830

**Galtier N, Gouy M, Gautier C** (1996) SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in the Biosciences* **12**, 543-548

**Glazer AN, Kechris KJ** (2009) Conserved amino acid sequence features in the alpha subunits of MoFe, VFe, and FeFe nitrogenases. *PLoS One* **4**, e6136

**Garza-Garcia A, Harris R, Esposito D, Gates PB, Driscoll PC** (2009) Solution structure and phylogenetics of prod1, a member of the three-finger protein superfamily implicated in salamander limb regeneration. *PLoS One* **4**, e7123

**Gonnet GH, Cohen MA, Benner SA** (1992) Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443-1445

**Gotoh O** (1993) Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Bioinformatics* **9**, 361-370

**Gyulai G, Humphreys H, Lagler R, Szabó Z, Tóth Z, Bittsánszky A, Gyulai F, Heszky L** (2006) Seed remains of common millet from the 4th (Mongolia) and 15th (Hungary) centuries: AFLP, SSR and mtDNA sequence recoveries. *Seed Science Research* **16**, 179-191

**Hahn Y** (2009) Molecular evolution of TEPP protein genes in metazoans. *Biochemical Genetics* **47**, 651-664

**Henikoff S, Henikoff JG** (1992) Amino acid substitution matrices from protein blocks. *Proceeding of the National Academy of Sciences USA* **89**, 10915-10919

**Higgens DG, Taylor WR** (2000) Multiple sequence alignment. In: *Protein Structure Prediction, Methods in Molecular Biology* (Vol 143), Humana Press, NY, pp 1064-3745

**Huang X, Madan A** (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877

**Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ** (2008) The 20 years of PROSITE. *Nucleic Acids Research* **36**, D245-249

**Kalsum HU, Shah ZA, Othman RM, Hassan R, Rahim SM, Asmuni H, Taliba J, Zakaria Z** (2009) SPlitSSI-SVM: An algorithm to reduce the misleading and increase the strength of domain signal. *Computational Biology* **39**, 1013-1019

**Kalendar R** (2009) FastPCR software for PCR primer and probe design and repeat search. Available online: www.biocenter.helsinki.fi/bi/programs/fastpcr.htm

**Kocabıyık S, Demirok B** (2009) Cloning and overexpression of a thermostable signal peptide peptidase (SppA) from *Thermoplasma volcanium* GSS1 in *E. coli*. *Biotechnology Journal* **4**, 1055-1065

**Kumar S, Dudley J, Nei M, Tamura K** (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics* **9**, 299-306

**Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG** (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948

**Le SQ, Gascuel O** (2008) An improved general amino acid replacement matrix. *Molecular Biology Evolution* **25**, 1307-1320

**Lipman DJ, Altschul SF, Kececioglu JD** (1989) A tool for multiple sequence alignment. *Proceeding of the National Academy of Sciences USA* **86**, 4412-4415

**Madhusudhan MS, Webb BM, Marti-Renom MA, Eswar N, Sali A** (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Engineering, Design and Selection* **22**, 569-574

**Mindnich RD, Penning TM** (2009) Aldo-keto reductase (AKR) superfamily: genomics and annotation. *Human Genomics* **3**, 362-370

**Moretti S, Reinier F, Poirot O, Armougom F, Audic S, Keduas V, Notredame C** (2006) PROTOGENE: turning amino acid alignments into *bona fide* CDS nucleotide alignments. *Nucleic Acids Research* **34**, W600-3

**Moretti S, Armougom F, Wallace IM, Higgins DG, Jongeneel CV, Notredame C** (2007) The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Research* **35**, W645-8

**Pellionisz A** (2008) The principle of recursive genome function. *Cerebellum* **7**, 348-359

**Poirot O, O'Toole E, Notredame C** (2003) Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Research* **31**, 3503-3506

**Rausch T, Emde AK, Weese D, Döring A, Notredame C, Reinert K** (2008) Segment-based multiple sequence alignment. *Bioinformatics* **24**, 187-192

**Shaik A, Khan M, Jamil K** (2009) Phylogenetic analysis of ALAD and MGP genes related to lead toxicity. *Toxicology and Industrial Health* **25**, 403-409

**Shui QY** (2008) *Bioinformatics: A Practical Approach* (1st Edn) Mathematical and Computational Biology Series, Chapman & Hall CRC, London, UK, 618 pp

**Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G** (2008) BLOSUM62 miscalculations improve search performance. *Nature Biotechnology* **26**, 274-275

**Suyama M, Torrents D, Bork P** (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**, W609-W612

**Szabó Z, Gyulai G, Humphreys M, Horváth L, Bittsánszky A, Lágler R, Heszky L** (2005) Genetic variation of melon (*C. melo*) compared to an extinct landrace from the Middle Ages (Hungary) I. rDNA, SSR and SNP analysis of 47 cultivars. *Euphytica* **146**, 87-94

**Tamura K, Dudley J, Nei M, Kumar S** (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599 (http://www.megasoftware.net)

**Taylor WR** (1988) A flexible method to align large numbers of biological sequences. *Journal of Molecular Evolution* **28**, 161-169

**Taylor WR, Flores TP, Orengo CA** (1994) Multiple protein structure alignment. *Protein Science* **3**, 1858-1870

**Taylor WR, Hatrick K** (1994) Compensating changes in protein multiple sequence alignments. *Protein Engineering* **7**, 341-348

**Thompson JD, Plewniak F, Poch O** (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**, 87-88

**Tóth Z, G Gyulai, Z Szabó, F Gyulai, L Heszky** (2008) New *Citrullus* haplotypes at the *tRNA*-Val – *rps*12 locus of cpDNA. In: Pitrat M (Ed) *Cucurbitaceae 2008, Proceedings of the IXth EUCARPIA Meeting on Genetics and Breeding of Cucurbitaceae*, May 21-24, INRA, Avignon, France, pp 335-340

**Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ** (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191

**Yap IV, Nelson RJ** (1996) Winboot: a program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms. IRRI, Manila