Published online 11 October 2012 *Nucleic Acids Research, 2013, Vol. 41, No. 1* ***e26***
doi:10.1093/nar/gks919

Predicting the accuracy of multiple sequence alignment algorithms by using computational intelligent techniques

Francisco M. Ortuño[1,*], Olga Valenzuela[2], Hector Pomares[1], Fernando Rojas[1], Javier P. Florido[3], Jose M. Urquiza[4] and Ignacio Rojas[1]

[1]Department of Computer Architecture and Computer Technology, [2]Department of Applied Mathematics, University of Granada (UGR), 18071 Granada, [3]Medical Genome Project, Andalusian Human Genome Sequencing Centre (CASEGH), 41092 Seville and [4]Chromatin and Disease Group, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet, Barcelona 08907, Spain

Received April 16, 2012; Accepted September 11, 2012

ABSTRACT

Multiple sequence alignments (MSAs) have become one of the most studied approaches in bioinformatics to perform other outstanding tasks such as structure prediction, biological function analysis or next-generation sequencing. However, current MSA algorithms do not always provide consistent solutions, since alignments become increasingly difficult when dealing with low similarity sequences. As widely known, these algorithms directly depend on specific features of the sequences, causing relevant influence on the alignment accuracy. Many MSA tools have been recently designed but it is not possible to know in advance which one is the most suitable for a particular set of sequences. In this work, we analyze some of the most used algorithms presented in the bibliography and their dependences on several features. A novel intelligent algorithm based on least square support vector machine is then developed to predict how accurate each alignment could be, depending on its analyzed features. This algorithm is performed with a dataset of 2180 MSAs. The proposed system first estimates the accuracy of possible alignments. The most promising methodologies are then selected in order to align each set of sequences. Since only one selected algorithm is run, the computational time is not excessively increased.

INTRODUCTION

Multiple sequence alignment (MSA) is a widely used approach in the current molecular biology. This technique involves the comparison of new sequences with well-known ones, extracting their shared information and their significant differences (1). MSA methods have traditionally been essential for analyzing biological sequences and designing applications in structural modeling, functional prediction, phylogenetic analysis and sequence database searching (2). Current MSA tools are also applied to comparisons of protein structures (3), reconstructions of phylogenetic trees (4) or predictions of mutations (5) and interactions (6).

More recently, the interest of MSA methodologies has even increased due to new experimental techniques. Current technologies provide a large amount of data that must be analyzed, processed and assessed. Consequently, new computational strategies were required to extract biological meanings from such information. Thus, supervised learning algorithms have been widely implemented in the analysis of genomic and proteomic experimental data. Additionally, recent experimental methods also retrieve further biological data, which is useful for extending the information included within alignment methods. Thus, current MSAs tools take advantage of heterogeneous features, which are provided by recent biological progress in functional, structural and genomic researches, to obtain more accurate alignments within a reasonable time (7). Therefore, MSAs are becoming one of the more powerful and essential procedures of analysis (8).

Traditionally, alignment strategies are mainly incorporated in progressive algorithm and consistency-based methods (7). Progressive algorithms assemble previously built pairwise alignments through a clustering method and store their evaluations in a library. Some well-known programs using progressive strategies are ClustalW (9) or Muscle (10). On the other hand, consistency-based methodologies, e.g. T-Coffee (11) or MSAProbs (12), develop consistency scoring schemes, taking into

*To whom correspondence should be addressed. Tel: +34 958 241778; Fax: +34 958 248993; Email: fortuno@atc.ugr.es

consideration not only the previous pairwise alignments but also if these alignments are consistent with the final result. However, neither progressive nor consistency-based methods build optimal alignments when sequences are distantly related. More recent approaches, such as 3D-COFFEE (13) or Promals (14), include further information (structure, domains or homologies) in addition to the provided sequences. Such features are usually found by experimental resources in databases such as Protein Data Bank (PDB) (15), Uniprot (16) or Pfam (17). Nevertheless, the consumed time is still excessive for these strategies and improvements are only relevant when sequences are evolutionarily less related (7).

Therefore, currently there are many alignment methodologies based on different strategies. Moreover, each MSA tool usually depends on particular features; thereby, there is no consensus about which one produces more accurate alignments (18,19). A new intelligent algorithm based on least square support vector machine (LS-SVM) is proposed here in order to predict how accurately each MSA tool will align a set of sequences. Interesting features related to the sequences and their products have been added from several resources in order to make this prediction. To the best of our knowledge, there are no similar studies in the current bibliography, which address the prediction of the alignment accuracy. This algorithm also estimates which methodologies are more significant to align those sequences. The system has been created from 218 sets of sequences provided by the BAliBASE benchmark (20) and their corresponding features. Since our algorithm applies a priori features to predict the accuracy, only the best method is run. Consequently, the CPU cost is not excessively increased.

MATERIALS AND METHODS

In this work, a novel system called 'Prediction of Accuracy in Alignments based on Computational Intelligence' (PAcAlCI) is developed. PAcAlCI is composed by four independent modules (Figure 1). First, 218 groups of sequences are aligned through 10 different methodologies, producing a dataset of 2180 alignments ('Input Dataset' module). Alignments are then evaluated in order to measure their accuracies. From these groups of sequences, several features are also retrieved from various relevant databases ('Feature Extraction' module). The most useful features are progressively included in a subset which is used by the subsequent algorithm ('Feature Selection' module). Finally, selected features are added to an LS-SVM model to predict alignment accuracies and, subsequently, the most suitable methodologies ('LS-SVM Prediction' module). The PAcAlCI system was completely implemented with Matlab (Version R2010b). The source code is available at http://www.ugr.es/~fortuno/PAcAlCI/PAcAlCI.zip.

Input dataset

A set of sequences must be considered in order to compare different alignment algorithms and develop the proposed prediction. Several datasets and techniques have usually been developed to standardize the comparison of alignment results, e.g. Oxbench (21), HOMSTRAD (22), Prefab (10) or BAliBASE (20). In this work, the BAliBASE benchmark (v3.0) was chosen.

BAliBASE defines several groups of sequences that can be easily aligned by standard algorithms. This dataset includes a total of 218 sets of sequences that were manually extracted from different databases, specifically the PDB (15). This benchmark also provides a set of handmade reference alignments (gold standard) in order to compare them with the alignments obtained by other tools. Thus, BAliBASE calculates a Sum-of-Pairs (SP) score to evaluate such alignments. These SP scores are used by our system to measure the quality of each alignment.

Sequences in BAliBASE are classified in the next subsets: (i) equidistant PDB sequences with <35 insertions and sharing <20% of identity between any pair of sequences (*RV11* and *RV12* subsets); (ii) PDB orphan sequences of families with >40% of identity and at least one known 3D structure (*RV20* subset); (iii) subfamilies of sequences that share >40% of identity but <20% with other subfamilies (*RV30* subset); (iv) sequences with >20% of identity and large terminal extensions (*RV40* subset) and (v) sequences with >20% of identity and internal insertions (*RV50* subset). From these subsets, it will be possible to establish, for example, which methodology is better when less related sequences are aligned or what differences are found when the methodologies

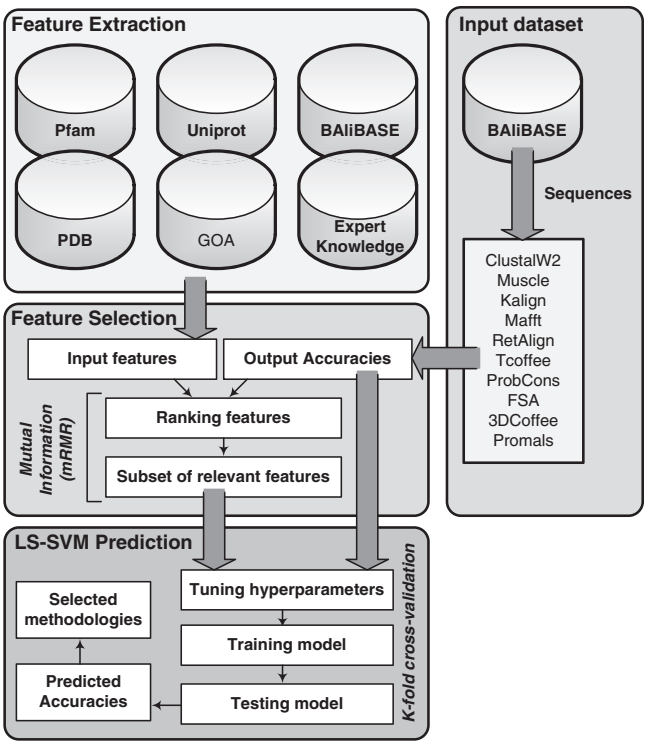

Figure 1. PAcAlCI scheme. The architecture is developed into four modules: input dataset, feature extraction, feature selection and LS-SVM prediction.

include additional information. These questions will be solved in the 'Comparison of MSA Methodologies' section.

MSA methodologies

Ten of the most relevant MSA tools are selected to be included in PAcAlCI. These tools are classified according to their implemented strategy: progressive techniques, consistency-based methods or algorithms including additional information (see summary in Table 1). Programs were run with their default features. Among the progressive methods, ClustalW (9), Muscle (10), Kalign (23), Mafft (24) and RetAlign (25) were chosen. ClustalW designs a tree-computing algorithm to find the alignment by means of distance scores and a gap weighting scheme. Muscle develops a strategy based on three stages, where a quickly built alignment is refined with an iterative method and a tree-dependent partitioning approach. Kalign uses the Wu–Manber string-matching algorithm (26) to improve the distance calculation of the classical progressive approach. Mafft identifies common homologies in sequences through a fast Fourier transform, significantly reducing the computational cost. Lastly, RetAlign implements a progressive corner-cutting algorithm to identify optimal alignments in a network of possible alignments.

Other three consistency-based approaches were included in PAcAlCI: T-Coffee (11), ProbCons (27) and Fast Statistical Alignment (FSA) (28). T-Coffee develops a standard consistency algorithm, building pairwise alignments and evaluating them against third sequences. ProbCons defines a probabilistic consistency based on a pair of hidden Markov models (pair-HMMs) to perform a novel scoring scheme for the standard consistency library. FSA estimates the insertion and deletion processes in sequences through pair-HMMs to combine their probabilities into alignments.

Finally, two more complex methodologies, namely 3D-Coffee (13) and Promals (14) were also applied. 3D-Coffee introduces structural information in the standard T-Coffee evaluations from the PDB (15), performing comparisons between each two structures and each sequence with its structure. On the other hand, Promals adds information based on homologies, combining sequences and homologies in profiles through HMMs.

Table 1. Summary of applied methodologies

Method	Version	Type
ClustalW (9)	2.0.10	Progressive
Muscle (10)	3.8.31	Progressive
Kalign (23)	2.04	Progressive
Mafft (24)	6.85	Progressive
RetAlign (25)	1.0	Progressive
T-Coffee (11)	8.97	Consistency-based
ProbCons (27)	1.12	Consistency-based
FSA (28)	1.15.5	Consistency-based
3D-Coffee (13)	8.97	Additional data
Promals (14)	vServer	Additional data

Ten different methodologies were run to align multiple sequences. Their versions and the applied strategies are also shown.

Databases and feature extraction

Features of BAliBASE sequences are extracted from well-known biological databases. Such databases are consulted to obtain interesting data which complement the sequences and to build a complete set of features. The final dataset will be composed by 23 features (see summary in Table 2).

Some features related to sequences, domains, amino acid types or structures have already been successfully included in other similar knowledge-based systems (18,29). However, the set of features is complemented with further information based on other studies such as protein interaction prediction (30) or protein model classification (31). Therefore, a more complete feature environment is presented in this work in order to study its relevance to sequence alignments. Here below, each consulted database is described, indicating which features have been retrieved and their nomenclature in the feature list:

- **BAliBASE** (20) can be considered the first consulted database, as it provides the sequences that are aligned. Then, the features associated with each set of sequences are the number of sequences (f_1), the average length of sequences (f_2) and the normalized

Table 2. Summary of features extracted from several databases

	Feature	Source	Range	Type	Rank
f_1	Number of sequences	BAliBASE	[4, 142]	Integer	3
f_2	Average length	BAliBASE	[66.13, 1630.11]	Real	4
f_3	Variance length (normalized)	BAliBASE	[0, 1]	Real	6
f_4	Reference subset	BAliBASE	[1, 6]	Integer	5
f_5	AA in α-helix[a]	UniProt	[0, 1]	Real	16
f_6	AA in β-strand[a]	UniProt	[0, 1]	Real	7
f_7	AA in transmembrane[a]	UniProt	[0, 1]	Real	22
f_8	Domains[b]	Pfam	[0.00,6.67]	Real	1
f_9	Shared Domains[b]	Pfam	[0.00,117.07]	Real	15
f_{10}	GO terms[b]	GOA	[0.00, 8.67]	Real	11
f_{11}	MF-GO terms[b]	GOA	[0.00, 5.17]	Real	17
f_{12}	CC-GO terms[b]	GOA	[0.00, 2.46]	Real	20
f_{13}	BP-GO terms[b]	GOA	[0.00, 4.07]	Real	19
f_{14}	Shared GO terms[b]	GOA	[0.00, 201.85]	Real	18
f_{15}	3D-Structures[b]	PDB	[0.04, 3.06]	Real	14
f_{16}	Seq. with any 3D structure	PDB	[0, 1]	Real	21
f_{17}	Shared 3D structures[b]	PDB	[0.00, 0.75]	Real	23
f_{18}	Polar AA[a]	Biochemistry	[0, 1]	Real	9
f_{19}	Non-polar AA[a]	Biochemistry	[0, 1]	Real	12
f_{20}	Basic AA[a]	Biochemistry	[0, 1]	Real	10
f_{21}	Aromatic AA[a]	Biochemistry	[0, 1]	Real	13
f_{22}	Acid AA[a]	Biochemistry	[0, 1]	Real	8
f_{23}	MSA method	—	[1, 10]	Integer	2

Twenty-three features were retrieved from different databases. The relevance ranking was also measured according to the mRMR procedure. [a]These features are calculated as the percentage of amino acids (AA) with that specific feature. [b]These features are calculated as the number of occurrences per sequence.

variance of the sequence length (f_3). This information determines whether there is any dependence between alignment tools and the number/length of sequences. Since BAliBASE classifies sequences according to certain features (see the 'Input Dataset' section for details), the subset, where each set of sequences is included, is proposed as another feature (f_4).

- **Uniprot** (16) consists of a wide repository of proteins with accurate, consistent and rich annotation. Several features are calculated from this database: the percentage of amino acids in α-helix structures (f_5), the percentage of amino acids in β-strand structures (f_6) and the percentage of amino acids in the transmembrane region (f_7). Data associated with similar secondary structures or locations usually indicate more related sequences or regions.
- **Pfam** (17) identifies common functional regions in families, also called domains. Domain features are performed from this database as: average number of domains per sequence (f_8) and average number of shared domains (between each pair of sequences) per sequence (f_9). Domain similarities usually imply regions with related functionality. This functionality can be useful to understand how some sequences must be efficiently aligned or how close sequences are in their families.
- **The Gene Ontology Annotation** (32) provides controlled vocabularies for the annotations of molecular attributes in different model organisms. These vocabularies are classified into three structured ontologies organized as a directed acyclic graph (DAG): molecular function (MF), cellular component (CC) and biological process (BP). Features used in this work from Gene Ontology Annotation (GOA) are average number of annotated terms per sequence (f_{10}); average number of annotated terms for each ontology per sequence: MF (f_{11}), CC (f_{12}) and BP (f_{13}) and average number of shared GO terms (between each pair of sequences) per sequence (f_{14}).
- **PDB** (15) includes information about experimentally determined 3D structures of each protein. The average number of annotated PDB structures per sequence (f_{15}), the percentage of sequences with structures (f_{16}) and the average number of shared structures (between each pair of sequences) per sequence (f_{17}) are proposed from this database.

Apart from these databases, other resources have been applied in order to complete the set of features. For instance, the classification of amino acids included in (33) has been applied to define: the percentage of polar uncharged amino acids [G,A,P,V,L,I,M] (f_{18}), the percentage of non-polar aliphatic amino acids [S,T,C,N,Q] (f_{19}), the percentage of basic positively charged amino acids [K,R,H] (f_{20}), the percentage of aromatic amino acids [F,W,Y] (f_{21}) and the percentage of negatively charged amino acids [D,E] (f_{22}).

Finally, the MSA method being executed (see the 'MSA methodologies' section) is the last included feature (f_{23}). This feature is determinant in the proposed system, as the purpose is to predict the accuracy of each method according to all these features. Besides, the most suitable methods according to the predicted accuracies are selected from that feature. Also, the accuracies of each method are included as outputs. As explained before, this accuracy is called SP score by BAliBASE and it is defined as a similarity value against the gold-standard references.

Feature selection based on mutual information

The relevance of the previously proposed features is analyzed through a feature selection procedure. Feature selection algorithms allow reducing the number of features, filtering out those irrelevant or redundant. One of the well-known feature selection, called minimal-redundancy-maximal-relevance (mRMR) (34), is applied in this work. First, this approach calculates the relevance of the features by using their mutual information. The obtained relevance is then assessed through the subsequent machine-learning procedure. The aim of mRMR is to select a feature at a time with a first-order incremental search, trying to avoid redundant features (see Supplementary mRMR feature selection for details). Discrete and continuous random variables are both considered in the mRMR algorithm. Such property is essential in the proposed set of features, since both types of variables were included ('real' and 'integer' types in Table 2). Besides, the output accuracy is also defined as a continuous variable.

The mRMR method achieves a great accuracy in a reduced time. Thus, the algorithm is useful to accurately select features among a huge number of them. The proposed features are then ranked from mRMR. Subsequently, the LS-SVM model is trained and evaluated progressively increasing the number of features from this ranking.

Least squares support vector machine

Features selected before are included in an LS-SVM model (35) in order to estimate different alignment accuracies. Subsequently, the algorithm also determines which tools are more likely to obtain the best alignment in term of accuracy. LS-SVMs models were generally designed for prediction approaches, but they present the most effective performance for regression problems. Since our system includes continuous values of accuracy, the proposed prediction is defined as a regression problem. Additionally, as the order of input data in LS-SVM is arbitrary, any change of the order would not affect the modeling result (35,36). Therefore, the application of LS-SVM would be an effective and faithful solution.

A kernel model must also be selected to correctly design the prediction method based on LS-SVM. The radial basis function (RBF) kernel was chosen to be applied in the proposed methodology. Additionally, LS-SVMs based on RBF kernels must also be performed by two kinds of hyper-parameters: the regulation parameter and the kernel parameters. These hyper-parameters were optimized by cross-validation in the proposed LS-SVM system (details about kernels and hyper-parameters in LS-SVMs are

provided in the Supplementary LS-SVM models). The LS-SVM algorithm is developed here from the Matlab toolbox found in (37).

In order to assess the LS-SVM model, a 10-fold cross-validation procedure is performed. This procedure randomly divides the complete dataset (2180 problems) into 10 subsets of 218 problems. Nine subsets are then applied to train the proposed system. The training procedure includes the most relevant features and the posterior accuracy for each problem in the subset. Thus, hyper-parameters are tuned and the LS-SVM model is estimated. Subsequently, the last subset is used to test the estimated LS-SVM model. The accuracies from such subset are then predicted and compared with those already known. The training and test procedures are repeated 10 times with the 10 different subsets. The predicted accuracies are then validated by their errors against real ones. The prediction error is measured by means of the 'mean relative error' (MRE). Taking into account this error value, a confidence interval is proposed to select the most suitable methodologies. For a specific set of sequences, those methodologies whose accuracies exceed a confidence value (σ_s) are selected (see the MRE and σ_s equations in the Supplementary LS-SVM validation).

RESULTS AND DISCUSSION

Comparison of MSA methodologies

As described before, each MSA method proposes different solutions depending on certain conditions or features. For this reason, biologists and researchers do not agree with a generally accepted solution (19). Some methods have been developed to unify criteria and choose the most suitable alignment tool (21,22), but this is currently an open issue.

In order to understand the performance of MSAs, accuracies from several methodologies can be compared. Previous reviews (7,8) have already compared accuracies from BAliBASE subsets (SP scores) against the applied strategy (progressive, consistency-based or approaches with additional data). Generally, SP scores are quite similar independently of the methodologies. Only when more distant sequences are provided (<20% of identity), accuracies are significantly higher in methods including additional data. However, these strategies including additional data are clearly in disadvantage in terms of required time (7). Thus, we could suggest that, only in special cases with less related sequences, additional data are clearly useful.

This analysis supports the idea of using a system to previously decide which methodologies are most promising to obtain better alignments. Here, PAcAlCI predicts accuracies to decide whether differences are enough to select more sophisticated methods against faster ones. Therefore, this system not only predicts the most relevant methodologies, but it also estimates differences between alignment performances. We could then decide which method constructs an accurate enough alignment according to its predicted accuracy.

Selection of feature subset

The complete dataset was composed by 2180 different inputs. Such inputs were retrieved from the 218 groups of sequences of BAliBASE. They were then aligned by the 10 previously proposed algorithms. For each input alignment, a set of 23 features was also retrieved. Output values were represented by the 2180 accuracies calculated from the input alignments.

As described above, the mRMR algorithm (34) was applied to select significant features. That procedure returned a ranking of features according to their relevance against calculated accuracies. An increasingly higher subset of features was then included in the subsequent system. According to this ranking (Table 2), the most relevant features were 'the number of domains' (f_8) and 'the applied methodology' (f_{23}). Regarding the first one, domains can be considered a measure of how deeply sequences are known. Domains are also associated with functional relationships and they involve more conserved sequence sections. Then, sequences that include more number of domains will be harder to align and, subsequently, the system could provide accurate predictions. On the other hand, the second feature is an essential variable because it is including obligatory information. This feature must always be included in order to know for which methodology the prediction is done, developing a robust and coherent system of prediction.

The features related to sequences, 'the number of sequences' (f_1) and 'the average/variance of the length' (f_2,f_3), were also ranked among first positions in the ranking. These features are highlighted because the availability to obtain accurate alignments directly depends on sequence properties. Other features less related to sequences but including amino acid information were found in the first half of the ranking. Features such as 'types of amino acids' ($f_{18} - f_{22}$) or 'the secondary structure' (f_5,f_6) provide complementary information about the composition and formation of sequences. Thus, they can also be helpful to efficiently predict some similarities.

Additionally, it is also important to analyze the occurrences in BAliBASE of the selected features in order to understand the obtained feature selection. BAliBASE sequences usually have known secondary structures (α-helix or β-strand) or GO terms. However, PAcAlCI was also trained with a few datasets from BAliBASE where these features are not known; thereby, cases without this information were also considered. Thus, new datasets from users not including that information can also be accurately estimated, returning their predicted accuracies and a set of suitable methods to use. On the other hand, datasets associated with other features (e.g. transmembrane regions) are considerably less included in BAliBASE. Consequently, the significance obtained for such feature was considered irrelevant and it was discarded from the selection procedure (for example, the 'transmembrane amino acids' feature was ranked in the 22nd position).

Prediction of alignment accuracy

Features previously analyzed were added to the subsequent LS-SVM model. PAcAlCI then predicted the

accuracy which each methodology returned for a set of sequences. As far as we are concerned, similar accuracy predictions in MSAs have not been addressed before. PAcAlCI was performed using an incremental combination of features in ascendant relevance order. Such combination was applied adding a feature at a time according to the previous ranking. Finally, a 10-fold cross-validation was performed to assess the algorithm. The prediction error (MRE) was calculated for the training and test sets.

The evolution of the errors for each combination of features is shown in Figure 2. According to such evolution, the error progressively decreases with higher number of features. However, an almost optimal value is reached from around 10 features. The prediction error is then kept around 6% for the training data and 9% for the test data. So, we could suggest that all features are not necessary to obtain the optimal prediction. A smaller number of features was then used to perform the system without lack of accuracy. Specifically, the 10 most relevant features were added to the LS-SVM model. According to this configuration, accuracies predicted from four sets of sequences are shown in Table 3. The total MRE value returned by PAcAlCI was 0.0587 for the training set and 0.1012 for the test. This error is distributed along the 2180 predicted accuracies as shown in Figure 3.

Analyzing more deeply the proposed system, higher error values are less frequent and they are usually associated with low accuracies (see detail in Figure 3). Alignments with low accuracies are less meaningful in our system, as their performances are totally unacceptable. Consequently, they could not even be considered in PAcAlCI. A minimal accuracy value, called α, was then defined as a threshold. Thus, the LS-SVM model only kept the most accurate alignments, filtering out the remaining ones. This threshold allowed improving the subsequent prediction. For example, if $\alpha = 0.5$, the MRE value using 10 input features decreased to 0.0340 in the training set and to 0.0608 in the test. As appreciated, error values were reduced by >2% and >4% for the training and test set, respectively. These errors improved because low values of accuracy, which led to highly wrong predictions, were previously filtered (see the new distribution of errors in Figure 4). Prediction errors are now considered low enough to adequately determine differences between methodologies. Then, the most suitable alignments can be selected according to their predicted accuracies.

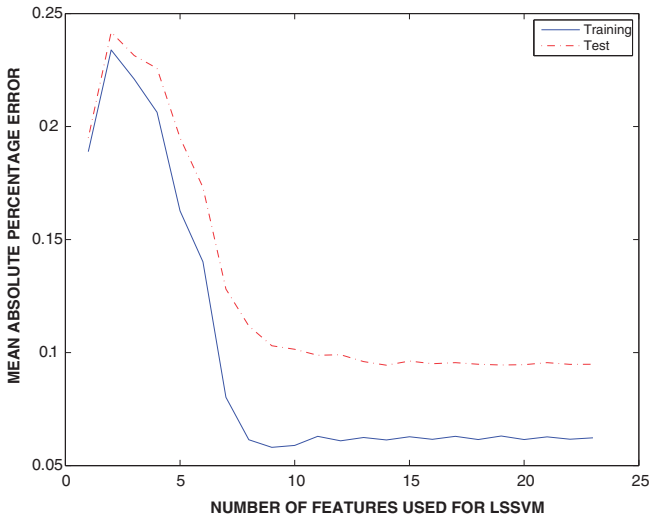

Figure 2. Evolution of the MRE. The number of features progressively increases in ascendant relevance order. Training and test errors are shown.

Selection of alignment methods

In occasions, several alignment methodologies obtain quite similar accuracies; thereby, no one significantly stands out from the rest. Consequently, as in other few researches (29,38), the most promising MSA tools can be selected according to several features. In this case, a confidence interval was defined to decide those methodologies which acceptably align a set of sequences. The confidence interval covers those accuracy values that

Table 3. Accuracies obtained for four different sets of sequences

Alignment	Method	Real Acc.	Pred. Acc.	Rel. error
RV11 4th	3D-Coffee	**0.7860**	**0.6478**	0.1758
	Promals	**0.7480**	**0.7068**	0.0551
	ProbCons	0.6230	**0.6836**	0.0973
	T-Coffee	0.6120	0.5976	0.0235
	Muscle	0.6000	0.3840	0.3600
	Kalign	0.5730	0.6168	0.0765
	Mafft	0.5260	0.6246	0.1875
	FSA	0.4390	0.4159	0.0527
	RetAlign	0.3880	0.2767	0.2868
	ClustalW2	0.1960	0.5291	1.6994
RV11 20th	3D-Coffee	**0.8540**	**0.7353**	0.1390
	Promals	**0.8170**	**0.8005**	0.0202
	Mafft	0.6920	0.6516	0.0583
	ProbCons	0.6810	0.6994	0.0270
	T-Coffee	0.6540	0.6035	0.0772
	ClustalW2	0.6520	0.5785	0.1127
	RetAlign	0.6330	0.5269	0.1677
	Kalign	0.6000	0.6823	0.1371
	Muscle	0.5920	0.6040	0.0203
	FSA	0.5320	0.6311	0.1863
RV40 24th	Promals	**0.6920**	**0.6259**	0.0955
	Mafft	**0.6760**	**0.6918**	0.0234
	Kalign	**0.6310**	0.5616	0.1099
	3D-Coffee	0.5750	0.6117	0.0638
	T-Coffee	0.5750	0.5562	0.0326
	ProbCons	0.5680	0.5982	0.0532
	FSA	0.5330	0.5331	0.0001
	Muscle	0.5140	0.5153	0.0024
	RetAlign	0.5110	0.5520	0.0802
	ClustalW2	0.4960	0.4378	0.1173
RV50 10th	Promals	**0.8650**	**0.7855**	0.0919
	Mafft	**0.7950**	**0.7536**	0.0520
	ProbCons	**0.7940**	**0.7547**	0.0495
	3D-Coffee	**0.7810**	**0.8055**	0.0314
	T-Coffee	**0.7790**	0.7105	0.0879
	Kalign	0.7370	**0.7496**	0.0171
	FSA	0.5910	0.6412	0.0849
	Muscle	0.5290	0.7089	0.3400
	RetAlign	0.5110	0.6308	0.2345
	ClustalW2	0.4830	0.5768	0.1942

Predicted accuracies are compared with those obtained by each methodology in four different problems. Values in bold show accuracies included in the confidence interval. The prediction error is also measured.

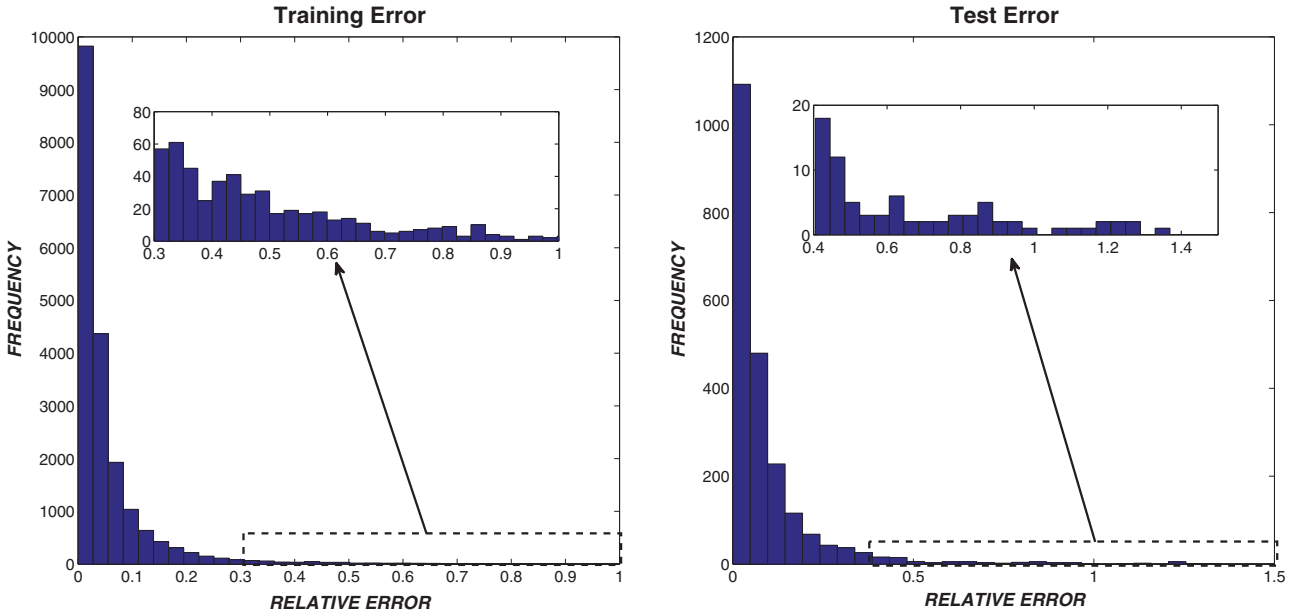

Figure 3. Distribution of relative errors for training and test sets. The corresponding LS-SVM prediction was performed using 10 features.

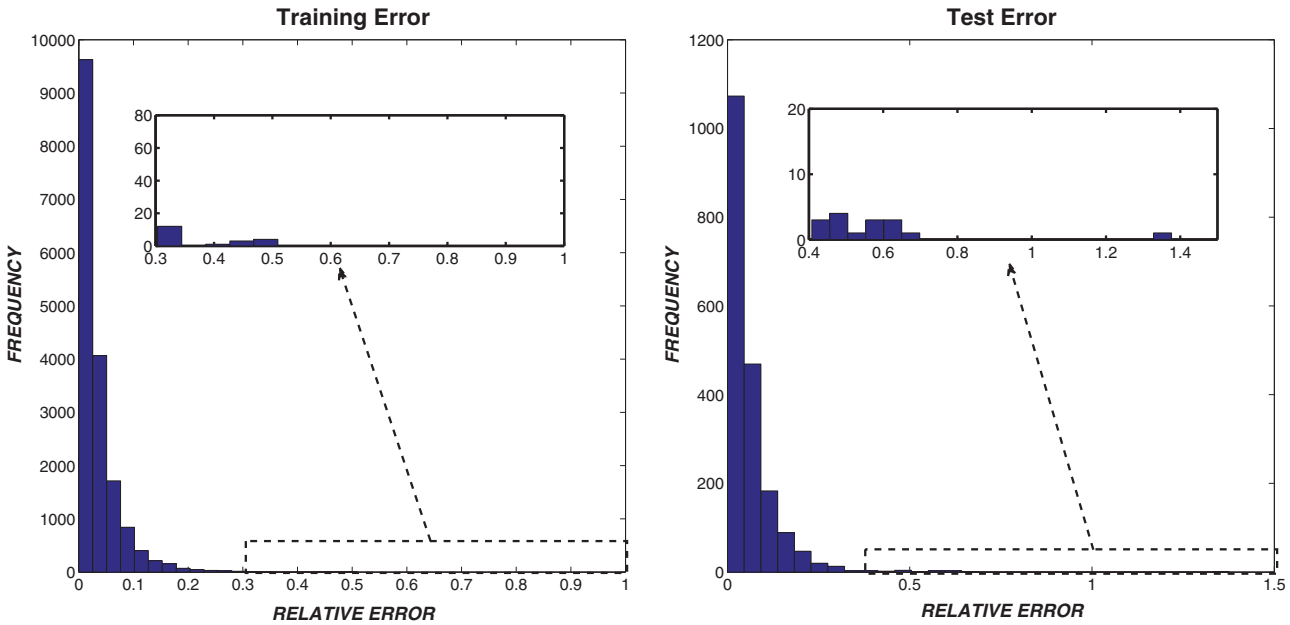

Figure 4. Distribution of relative errors for training and test sets. Low accuracies were previously filtered to improve the LS-SVM prediction, avoiding prediction with high errors ($\alpha = 0.5$).

are higher than a confidence value σ_s (see its formal definition in Supplementary LS-SVM validation). Those methodologies whose accuracies exceed such confidence value were chosen as candidate methods.

This confidence interval was applied to real accuracies and predicted accuracies. Two sets of suitable methodologies were then retrieved (real and predicted sets). Thus, the number of selected methodologies was variable for each group, as it depends on how similar accuracies were in that specific problem. Both groups were then compared in order to know how many methodologies were correctly selected in the predicted set (see four examples in Figure 5). For example, using accuracies from the performance of 10 features without α threshold, the 83.55% of predicted methodologies were also included in the real group. When accuracies were predicted with the limitation $\alpha = 0.5$, the percentage of successfully selected methodologies increased to 85.89%. Therefore, the proposed system usually performed an accurate group of outstanding methodologies.

As shown in the examples of Figure 5, methodologies including additional information, namely 3D-Coffee and Promals, were selected for sequences with low similarity (RV11 subset). In these cases, more commonly used

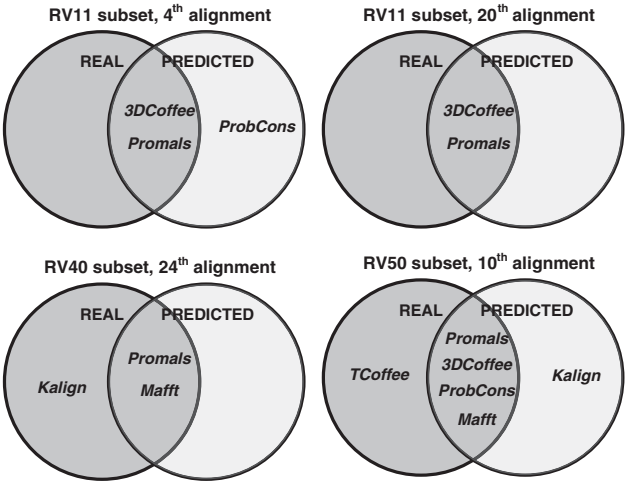

Figure 5. Intersection of suitable and predicted methodologies (Venn diagrams) corresponding to the four alignments whose accuracies are shown in Table 3.

aligners (ClustalW, Kalign or Muscle) were not selected, as they did not build accurate enough alignments. However, these more complex methods (3D-Coffee and Promals) did not significantly outperform other faster methods when sequences were more related. Thus, methodologies as Mafft, T-Coffee, Kalign or ProbCons were also selected when sequences were highly related (>20% of similarity). Consequently, we could again suggest that the prediction system is working as expected. For instance, those datasets including more than two domains per sequence selected Kalign as a suitable method (in the 80.95% of cases), whereas Mafft was appropriate for datasets with less domains (78% of datasets selected it). Regarding the size of the dataset, large datasets (>50 sequences or >400 amino acids of average length) usually picked both Mafft and Kalign (90.17 and 71.9%, respectively), while ProbCons was chosen for shorter datasets (62.89%). Finally, Kalign also suited when the sequence lengths in datasets have a high variability (a difference of >100 amino acids in average between sequence lengths) and ProbCons for low variability (69.05% and 65.89% of cases, respectively).

Although there are other expert systems to select adequate MSA tools (38,39), PAcAlCI was compared with AlexSys (29), as it performs a more similar strategy (see comparison in Table 4). However, comparing both methodologies can be complicated. Both systems develop similar machine-learning approaches, but their objectives are quite different. AlexSys defines a decision-tree approach to predict whether sequences are 'strongly' or 'weakly' aligned with each specific method (classification problem and binary solution). The best method among those classified as 'strong' is then inferred according to their success probability or their required CPU time. This binary classification can be quite subjective in some cases. Since accuracies over 0.5 are already classified as 'strong', quite different accuracies, e.g. 0.5 and 0.9, are considered identical in the AlexSys approach. In a different way, PAcAlCI first predicts accuracy values (regression problem and real solution). The accuracy prediction provides a relevant improvement in order to decide whether it is worth aligning with a specific methodology. Besides, suitable methods are also selected according to the best accuracies. In general, AlexSys correctly predicts the best aligner in a 45% of its test alignments. In another 45.5% of the alignments, the best aligner corresponded to the second predicted method. In general, global success rates in PAcAlCI are quite similar (83.55% or 85.89% depending on the α threshold), although the number of suitable methods is usually higher in our prediction. Regarding the included tools, PAcAlCI is composed by a wider group of previous methodologies (10 approaches compared with the six of AlexSys), including more complex ones as 3D-Coffee or Promals.

Table 4. Comparison between PAcAlCI and AlexSys

Feature	PAcAlCI	AlexSys
Number of aligners	10	6
Kind of problem	Regression (real)	Classification (binary)
Machine-learning strategy	LS-SVM	Decision trees
Values of prediction	Accuracies	Weak (Acc. < 0.5) Strong (Acc. > 0.5)
Success rate	83.6% ($\alpha = 0$) 85.9% ($\alpha = 0.5$)	45.0% (first aligner) 45.5% (second aligner)

PAcAlCI is qualitatively compared with AlexSys. The performance and attributes of both procedures are shown.

Despite these differences, both methods may be considered complementary, as both perform accurate classifiers but in different contexts. In any case, the final decision of selecting the most suitable methodology among the proposed ones can rely on the final user of this system. Other criteria such as the complexity of parameters in the methodologies or the required time could also be taken into account in order to choose the correct tool among the selected ones.

CONCLUSION

MSA is currently an open issue. Alignment tools must be continually improved, as they are essential in the analysis of huge amount of data provided by next-generation sequencing and high-throughput experiments. Thus, new trends in MSAs aim to integrate the major amount of information while trying to significantly reduce the used time. For this reason, efficient computational techniques are increasingly implemented.

In this work, a complete study of MSA methodologies has been developed. Relevant methodologies in this field were first compared. Several types of methods were discussed and we have suggested that only in special cases more sophisticated approaches including additional information are really necessary. A novel intelligent system (PAcAlCI) was then proposed based on the knowledge acquired from this study.

PAcAlCI was designed in order to predict the accuracy that each alignment method reaches for a specific set of

sequences. This information gives us an idea of how accurately each methodology works. The mRMR feature selection technique was first applied to 23 features previously retrieved from several biological databases. We have also described how the system can be performed with only the 10 most relevant features to predict accuracies with a reasonable efficiency. Finally, we have proposed the outstanding methodologies which can be used for certain sequences according to their predicted accuracies. In this sense, the proposed algorithm is able to successfully select the most outstanding methods according to the previously predicted accuracies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary mRMR feature selection, Supplementary LS-SVM models, Supplementary LS-SVM validation and Supplementary References [40–47].

FUNDING

The Spanish CICYT [Project SAF2010-20558]; the Government of Andalusia [Project P09-TIC-175476]. Funding for open access charge: the Government of Andalusia [Project P09-TIC-175476].

Conflict of interest statement. None declared.

REFERENCES

1. Attwood,T.K. and Parry-Smith,D.J. (2002) *Introduction to Bioinformatics*. Pearson Education, Prentice Hall.
2. Pei,J. (2008) Multiple protein sequence alignment. *Curr. Opin. Struct. Biol.*, **18**, 382–386.
3. Gelly,J.C., Joseph,A.P., Srinivasan,N. and de Brevern,A.G. (2011) iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res.*, **39**, W18–W23.
4. Wang,L.S., Leebens-Mack,J., Wall,P.K., Beckmann,K., dePamphilis,C.W. and Warnow,T. (2011) The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE–ACM Trans. Comput. Biol. Bioinform.*, **8**, 1108–1119.
5. Hicks,S., Wheeler,D.A., Plon,S.E. and Kimmel,M. (2011) Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum. Mutat.*, **32**, 661–668.
6. Li,A.X., Marz,M., Qin,J. and Reidys,C.M. (2011) RNA–RNA interaction prediction based on multiple sequence alignments. *Bioinformatics*, **27**, 456–463.
7. Kemena,C. and Notredame,C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, **25**, 2455–2465.
8. Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, **11**, 473–483.
9. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: improving the sensitivity of progressive multiple sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
10. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
11. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol Biol.*, **302**, 205–217.
12. Liu,Y., Schmidt,B. and Maskell,D.L. (2010) MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*, **26**, 1958–1964.
13. O'Sullivan,O., Suhre,K., Abergel,C., Higgins,D.G. and Notredame,C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.
14. Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
16. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H.Z., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
17. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
18. Nuin,P.A.S., Wang,Z.Z. and Elisabeth,R.M. (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471.
19. Sierk,M.L., Smoot,M.E., Bass,E.J. and Pearson,W.R. (2010) Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC Bioinformatics*, **11**, 146.
20. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
21. Raghava,G.P.S., Searle,S.M.J., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.
22. Stebbings,L.A. and Mizuguchi,K. (2004) HOMSTRAD: recent developments of the homologous protein structure alignment database. *Nucleic Acids Res.*, **32**, D203–D207.
23. Lassmann,T. and Sonnhammer,E.L. (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.
24. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
25. Szabo,A., Novak,A., Miklos,I. and Hein,J. (2010) Reticular alignment: a progressive corner-cutting method for multiple sequence alignment. *BMC Bioinformatics*, **11**, 570.
26. Wu,S. and Manber,U. (1992) Fast text searching allowing errors. *Commun. ACM*, **35**, 83–91.
27. Do,C.B., Mahabhashyam,M.S.P., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
28. Bradley,R.K., Roberts,A., Smoot,M., Juvekar,S., Do,J., Dewey,C., Holmes,I. and Pachter,L. (2009) Fast statistical alignment. *PLoS Comput. Biol.*, **5**, 5.
29. Aniba,M.R., Poch,O., Marchler-Bauer,A. and Thompson,J.D. (2010) AlexSys: a knowledge-based expert system for multiple sequence alignment construction and analysis. *Nucleic Acids Res.*, **38**, 6338–6349.
30. Wu,X.M., Zhu,L., Guo,J., Zhang,D.Y. and Lin,K. (2006) Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res.*, **34**, 2137–2150.
31. Roslan,R., Othman,R.M., Shah,Z.A., Kasim,S., Asmuni,H., Taliba,J., Hassan,R. and Zakaria,Z. (2010) Utilizing shared interacting domain patterns and Gene Ontology information to improve protein–protein interaction prediction. *Comput. Biol. Med.*, **40**, 555–564.
32. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
33. Mathews,C.K., Holde,K.E.V. and Ahern,K.G. (2000) *Biochemistry*. Benjamin Cummings Publishing, Redwood City, CA.

34. Peng,H., Long,F. and Ding,C. (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.
35. Suykens,J.A.K., Van Gestel,T., De Brabanter,J., De Moor,B. and Vandewalle,J. (2003) *Least Squares Support Vector Machines*. World Scientific Pub. Co. Inc., Singapore.
36. Li,L., Su,H. and Chu,J. (2009) Sparse representation based on projection method in online least squares support vector machines. *J. Control Theory Appl.*, **7**, 163–168.
37. De Brabanter,K., Karsmakers,P., Ojeda,F., Alzate,C., De Brabanter,J., Pelckmans,K., De Moor,B., Vandewalle,J. and Suykens,J.A.K. (2011), LS-SVMLab: a MATLAB toolbox for least squares support vector machines (v1.8). http://www.esat.kuleuven.ac.be/sista/lssvmlab (21 September 2012, date last accessed).
38. Anderson,C.L., Strope,C.L. and Moriyama,E.N. (2011) SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC Bioinformatics*, **12**, 184.
39. Thompson,J.D., Muller,A., Waterhouse,A., Procter,J., Barton,G.J., Plewniak,F. and Poch,O. (2006) MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*, **7**, 318.
40. Estevez,P.A., Tesmer,M., Perez,C.A. and Zurada,J.M. (2009) Normalized mutual information feature selection. *IEEE Trans. Neural Netw.*, **20**, 189–201.
41. John,G., Kohavi,R. and Pfleger,K. (1994) Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, 121–129.
42. Bins,J. and Draper,B.A. (2001) Feature selection from huge feature sets. In *8th IEEE International Conference on Computer Vision*, **2**, 159–165.
43. Cover,T.M. and Thomas,J.A. (2006) *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA.
44. Kullback,S. (1997) *Information Theory and Statistics*. Courier Dover Publications, New York, NY, USA.
45. Babich,G.A. and Camps,O.I. (1996) Weighted Parzen windows for pattern classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, **18**, 567–570.
46. Hestenes,M.R. and Stiefel,E. (1952) Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.*, **49**, 409–436.
47. Rossi,F., Lendasse,A., Francois,D., Wertz,V. and Verleysen,M. (2006) Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics Intell. Lab. Syst.*, **80**, 215–226.