BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 8 2012, pages 1078–1085
doi:10.1093/bioinformatics/bts098

Sequence analysis

Advance Access publication February 25, 2012

High-quality sequence clustering guided by network topology and multiple alignment likelihood

Vincent Miele*, Simon Penel, Vincent Daubin, Franck Picard, Daniel Kahn and Laurent Duret

Laboratoire Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, INRA, UMR5558, Villeurbanne, France

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Proteins can be naturally classified into families of homologous sequences that derive from a common ancestor. The comparison of homologous sequences and the analysis of their phylogenetic relationships provide useful information regarding the function and evolution of genes. One important difficulty of clustering methods is to distinguish highly divergent homologous sequences from sequences that only share partial homology due to evolution by protein domain rearrangements. Existing clustering methods require parameters that have to be set a priori. Given the variability in the evolution pattern among proteins, these parameters cannot be optimal for all gene families.

Results: We propose a strategy that aims at clustering sequences homologous over their entire length, and that takes into account the pattern of substitution specific to each gene family. Sequences are first all compared with each other and clustered into pre-families, based on pairwise similarity criteria, with permissive parameters to optimize sensitivity. Pre-families are then divided into homogeneous clusters, based on the topology of the similarity network. Finally, clusters are progressively merged into families, for which we compute multiple alignments, and we use a model selection technique to find the optimal tradeoff between the number of families and multiple alignment likelihood. To evaluate this method, called `HiFiX`, we analyzed simulated sequences and manually curated datasets. These tests showed that `HiFiX` is the only method robust to both sequence divergence and domain rearrangements. `HiFiX` is fast enough to be used on very large datasets.

Availability and implementation: The Python software `HiFiX` is freely available at `http://lbbe.univ-lyon1.fr/hifix`

Contact: vincent.miele@univ-lyon1.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 16, 2011; revised on February 20, 2012; accepted on February 23, 2012

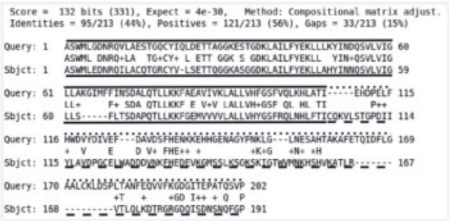
1 INTRODUCTION

Genomes are the result of a long evolutionary process that began >3 billion years ago. Reconstructing the evolutionary history of genes contained within these genomes is of major interest, not only to uncover the phylogeny of organisms, but also to understand the functioning of living systems. Thanks to the progress of genome sequencing projects, millions of protein-coding genes, from thousands of species, are now available in sequence databases. Several specialized databases have been developed with the aim of providing systematic information about the homology relationships between these sequences, either at the level of protein domains [such as ProDom (Bru *et al.*, 2005) or Pfam (Finn *et al.*, 2010)], or among entire proteins, considered as a whole [such as COG (Tatusov *et al.*, 2001), Treefam (Ruan *et al.*, 2008), EnsEMBL Compara (Vilella *et al.*, 2009) or HOGENOM (Penel *et al.*, 2009)]. The systematic analysis of homology relationships among sequences typically involves two steps: (i) pairwise comparison of all proteins to detect homology (e.g. with BLAST) and (ii) clustering of homologous proteins (or protein domains) into families. The choice of the clustering strategy is crucial for all subsequent analyses, and depends on the purpose of the study. Here, we will discuss specifically clustering strategies that aim at describing homology relationships between entire proteins, not protein domains.

By definition, a family of homologous sequences is a set of sequences that derive from a common ancestor. Hence, in principle, modular proteins containing domains with distinct evolutionary histories should not be included in the same family. The main problem is that the detection of homology by sequence similarity search is subject both to false positive (FP) and false negative (FN) errors. Typically, homology cannot be detected by sequence similarity when proteins have diverged too much. Given that the rate of evolution often varies along proteins, some sequences that are homologous over their entire length may be only locally alignable. Such cases would be erroneously considered as partially homologous, and will lead to FNs. Conversely, it has been shown that similarity search programs, such as BLAST, sometimes tend to extend the local alignment beyond the actual homologous domain (Gonzalez and Pearson, 2010). Thus, in some cases, modular proteins sharing only partial homology can be aligned over their entire length, and hence be classified in the same family (FPs, see Fig. 1). The rate of FPs can be decreased by using more stringent sequence similarity criteria, but this necessarily leads to an increase in the rate of FNs.

All clustering methods have some parameters that can be tuned to optimize the tradeoff between sensitivity and specificity. However, the choice of these parameters is totally empirical, and the default parameters that are proposed by the authors generally reflect the result of trial and error on a limited benchmark set of sequences (Apeltsin *et al.*, 2011; Wittkop *et al.*, 2010). One other problem

*To whom correspondence should be addressed.



```
Score =  132 bits (331), Expect = 4e-30,   Method: Compositional matrix adjust.
Identities = 95/213 (44%), Positives = 121/213 (56%), Gaps = 33/213 (15%)

Query: 1   ASWMLGDNRQVLAESTGQCYIQLDETTAGGKESTGDKLAILFYEKLLLKYINDQSVLVIG 60
            ASWML DNRQ+LA  TG+CY+ L ETT GGK S GDKLAILFYEKLL  YIN+QSVLVIG
Sbjct: 1   ASWMLEDNRQILACQTGRCYV-LSETTQGGKASGGDKLAILFYEKLLAHYINNQSVLVIG 59

Query: 61  LLAKGIMFFINSDALQTLLKKFAEAVIVKLALLVHFGSFVQLKHLATI-----EHDPELF 115
            LL+     F+ SDA QTLLKKF E V+V LALLVH+GSF QL HL TI        P++
Sbjct: 60  LLS-----FLTSDAPQTLLKKFGEMVVVVLALLVHYGSFRQLNHLFTICQKVLSTGPDII 114

Query: 116 HWDVYDIVEF---DAVDSFHENKKEHHGENAGYPNKLG---LNESAHTAKAFETQIDFLG 169
            +  V    E    D V+ FHE++ +            +K+G   +N+ +H
Sbjct: 115 YLAVDPGCELWADDDVNKEHEDEVKGMSSLKSGKSKIGTWMMNKHSHVKATLR------- 167

Query: 170 AALCKLDSPLTANFEQVVFKGDGITEPATQSVP 202
                     +T   +    +GD I++ + Q  P
Sbjct: 168 ---------VTLQLKDTRGRGDQISDNSNQFGP 191
```

Fig. 1. Example of over-extension of BLAST local alignment. The two protein sequences (Query and Sbjct) were obtained by simulation of sequence evolution performed with INDELible (Fletcher and Yang, 2009). Each protein consists of two domains of equal size: the N-terminal domain (plain line), which is homologous between both proteins, and a specific, non-homologous C-terminal domain (in dotted and dashed line, respectively). Although these proteins share homology only in the N-terminal domain, the BLAST alignment extends over the entire length of both proteins. Such an overextension can also be found with natural protein sequences (Gonzalez and Pearson, 2010).

is that the tempo of sequence evolution is highly variable among proteins, and hence optimal parameters may vary among families. Moreover, the optimal parameters may also vary according to the size of the sequence dataset. For example, we recently developed a clustering software [SiLiX, (Miele *et al.*, 2011)], based on a sophisticated divide-and-conquer procedure, which presents the advantage of being extremely fast and memory efficient, and hence can be used on very large sequence datasets (contrarily to most other existing methods). Benchmark tests showed that on average, SiLiX performs as well (or even better) than other methods in term of clustering quality (Miele *et al.*, 2011). However, given that SiLiX is based on transitive clustering, this method is expected to lead to FPs among large protein families, because the risk of illegitimately grouping gene families increases with the number of sequences.

To circumvent these problems, we propose here a new clustering strategy. The main idea is that the decision to include or not a protein in a given family should be based on the examination of multiple sequence alignments, not simply on the analysis of pairwise sequence similarities. Indeed, a multiple alignment contains information about the mode and tempo of evolution at each amino-acid position of a given protein family, which can be used to decide whether or not a new sequence belongs to that family. The strategy we propose consists in three steps:

(1) Rapid clustering with SiLiX, using low-stringency criteria, to get a first set of *pre-families*, with a low rate of FNs (but possibly including FPs).

(2) Decomposition of each pre-family into homogeneous protein clusters, by analysing the topology of similarity networks.

(3) Hierarchical clustering of previous clusters into families, by progressive multiple alignment of protein clusters and evaluation of alignment quality at each step.

In this article, we describe in details the methods used in the last two steps. We tested this procedure on several well-studied sets of biological sequences, and on a set of simulated sequences. These tests show a significant improvement over existing methods, notably for large and highly divergent protein families. The method, called HiFiX, is fast enough to be used in practice for very large datasets. HiFiX is available as a Python software at http://lbbe.univ-lyon1.fr/hifix.

2 METHODS

Similarity relationships can be modeled through graph theory representation, with sequences as vertices and similarities as edges forming a *similarity network*. Analyzing the topology of this network was proved to be relevant for studying protein or domain homology (Medini *et al.*, 2006; Song *et al.*, 2008), performing phylogenetic inference (Andrade *et al.*, 2011; Atkinson *et al.*, 2009; Zhang *et al.*, 2011) or improving homology measurement (Fokkens *et al.*, 2010). In the following, we propose to consider the topology as a guide, first to subdivide each pre-family into homogeneous sequence clusters, and then for the progressive merging of clusters into families.

2.1 Search for homogeneous clusters of homologous sequences as communities in networks

For the sake of simplicity, we now consider a single pre-family retrieved by SiLiX but it is straightforward to independently apply the following approach to all the pre-families. We consider the list of pairs of similar sequences that satisfy alignment coverage constraints (Miele *et al.*, 2011; Penel *et al.*, 2009). We build the similarity network G where the n vertices are sequences and the m edges correspond to the binary information determined by SiLiX of being similar sequences. G is connected by construction. We empirically observed that the network topology can display inhomogeneities (Fig. 2), with concentrations of edges varying between groups of vertices. Indeed, some similarity networks display a *community structure*, that is an organization of vertices in clusters, with many edges between vertices of the same cluster and significantly fewer edges between vertices from different clusters (Fortunato, 2010; Girvan and Newman, 2002).

At this point, the use of statistical approaches offers a powerful way to characterize these complex structural patterns present in networks. In particular, algorithms relying on the maximization of the *modularity* are very popular to decipher community structure (Blondel *et al.*, 2008; Girvan and Newman, 2002): given that the degree of a vertex is its number of incident edges, the modularity measures the difference between the observed fraction of edges inside clusters and the expected fraction for a random graph with the same degree distribution. To find communities in similarity networks, we chose the Louvain algorithm (Blondel *et al.*, 2008) that is documented to perform well and avoid resolution problems (Fortunato, 2010).

Similarly to previous studies (Andrade *et al.*, 2011; Medini *et al.*, 2006), the resulting communities correspond to homogeneous clusters of homologous sequences. However, we observed that homologous sequences belonging to a same protein family are also found in distinct communities, so that it is necessary to merge them into larger clusters that we call *meta-communities*.

2.2 Hierarchical clustering into meta-communities and model selection

Using the notations of Stochastic Block Models (Nowicki and Snijders, 2001; Picard *et al.*, 2009), the n vertices of G are distributed into Q clusters in proportion $\alpha = (\alpha_1, ..., \alpha_Q)$ and size $(n_1, ..., n_Q)$, such that $n_q = n\alpha_q$. We define the *label matrix* Z such that $Z_{iq} = 1$ if vertex i belongs to cluster q, 0 otherwise. The *connectivity matrix* π contains the probabilities π_{ql} that a vertex of cluster q is connected to a vertex of cluster l. In our case, we will not estimate Z using classical model-based clustering algorithms but using Louvain and the following hierarchical algorithm.

Starting from the K communities identified by Louvain, we iteratively merge sequences into meta-communities until all n sequences are in a single cluster. We use the π probabilities as a guide for hierarchical clustering so that we merge clusters of maximal connectivity. In this way, we obtain a hierarchy of clustering results of $Q = K$ to 1 clusters or meta-communities

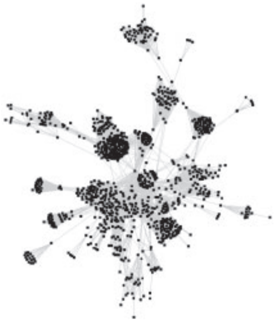


Fig. 2. Example of a community structure in the similarity network corresponding to HBG739407 HOGENOM family, retrieved by `SiLiX`. Network was drawn with Cytoscape (Shannon *et al.*, 2003).

(Algorithm 1). We denote by $\widehat{Z}$ the labels resulting from the hierarchical algorithm.

It is now necessary to apply a model selection technique to choose the clustering that is the most relevant for our purpose, i.e. that is most likely to define independent clusters of proteins that conform to uniform sequence models. For this purpose, we evaluate each clustering on the basis of alignment likelihood using profile-HMM models. We consider the set of sequences $S=\{S_1,...,S_n\}$ distributed into Q clusters, and we perform Q multiple alignments to derive the corresponding profile-HMM models (Durbin *et al.*, 1998) with parameters $\theta=(\theta_1,...,\theta_Q)$. We score the quality of the within cluster alignment using the following completed log-likelihood:

$$\log\mathcal{L}_Q(S,\widehat{Z};\alpha,\theta)=\underbrace{\log\mathcal{L}_Q(\widehat{Z};\alpha)}_{\sum_q n_q\log(\alpha_q)}+\log\mathcal{L}_Q(S|\widehat{Z};\theta), \qquad (1)$$

where

$$\log\mathcal{L}_Q(S|\widehat{Z};\theta)=\sum_{i,q}\widehat{Z}_{iq}\log\mathcal{L}_Q(S_i|\widehat{Z}_{iq};\theta_q). \qquad (2)$$

This corresponds to the log-likelihood of within cluster profile-HMMs for given clusters ($\widehat{Z}$) that accounts for cluster size heterogeneities [through term $\log\mathcal{L}_Q(\widehat{Z};\alpha)$]. Since we need a measure of the trade-off between alignment quality and number of clusters, we propose to use a penalized version of this joint log-likelihood by resorting to the *Integrated Classification Likelihood* criterion [ICL, (Biernacki *et al.*, 2000)]. Thus we choose the best number Q^* of clusters by maximizing:

$$\mathrm{ICL}(Q)=\log\mathcal{L}_Q(S,\widehat{Z};\widehat{\alpha},\widehat{\theta})-\left(|\widehat{\alpha}|+|\widehat{\theta}|\right)\frac{\log n}{2} \qquad (3)$$

with $|\widehat{\alpha}|=Q-1$ and $|\widehat{\theta}|$ being the number of free parameters for mixture proportions and profile-HMMs, respectively. At this final step, the Q^* meta-communities will define the Q^* families of homologous proteins. The whole procedure, called `HiFiX` for *HIgh FIdelity Clustering of Sequences*, is summarized in Algorithm 1.

Algorithm 1 `HiFiX` — *HIgh FIdelity Clustering of Sequences*

```
Input: $K$ clusters as communities
 1. perform $K$ multiple alignments
 2. build profile-HMM models $\mathcal{M}_q, 1 \le q \le K$
 3. compute ICL($K$)
 4. for $Q=K-1$ to 1 do
 5.     merge two clusters $q$ and $l$ with highest $\pi_{ql}$
 6.     perform alignment of alignments $q$ and $l$
 7.     build profile-HMM models $\mathcal{M}_q, 1 \le q \le Q$
 8.     calculate ICL($Q$)
 9.     update $\pi$
10. end for
Output: $Q^* = \mathrm{argmax}_Q \mathrm{ICL}(Q)$
```

Because π is used as a guide in the algorithm but not directly related to likelihoods, we add the possibility to examine the $I>1$ highest values of π_{ql} at line 5 of Algorithm 1. Following the strategy presented in (Han *et al.*, 2008), we select the merging that minimizes the log-likelihood loss per sequence.

2.3 The `HiFiX` software package

The presented algorithm is implemented in the `HiFiX` software program, written in `Python` and designed to be executed in parallel on multiprocessor architectures. In `HiFiX`, community detection is performed with the `Louvain` algorithm, available as a software package from `http://sites.google.com/site/findcommunities/` (Blondel *et al.*, 2008). Multiple alignments are computed with `mafft` and `mafft-profile` [`v6.849`, (Katoh *et al.*, 2009)] with default parameters. Log-likelihoods in Formula (2) are derived from Hidden Markov Models (HMMs) using `HMMER v3.0` (Eddy, 2009) as follows: (i) HMMs are constructed using `hmmbuild` with the `–enone` option in order to avoid rescaling of log-likelihood ratios; (ii) HMMER scores of all sequences in each community are obtained with `hmmsearch` and summed to give log-likelihood ratios; and (iii) changes in log-likelihood ratios are taken as changes in log-likelihoods because the background model is constant. Finally, at each iteration of Algorithm 1, we probe the $I=2$ highest values of π_{ql}.

`HiFiX` requires the use of `SiLiX` [(Miele *et al.*, 2011), default parameters recommended] as a preliminary step. `HiFiX` is licensed under the General Public License `http://www.gnu.org/licenses/licenses.html`.

2.4 Other program parameters

To perform our experiments, we ran BLASTP with the following options: `-M BLOSUM62 -G 11 -E 1 -e 1e-04 -v 10000 -b 10000 -F "m S" -m 8 -z 300000000.`

`MCL` (Enright *et al.*, 2002) was used with default parameters. `TransClust` threshold parameter was set at 55 as suggested by Wittkop *et al.* (2010). Maximal cluster size for `hcluster_sg` was set as the total number of proteins in each family.

We performed simulation of protein families using the `INDELible` program (Fletcher and Yang, 2009) with default settings, the WAG substitution model, a continuous gamma rate heterogeneity with $\alpha=1$ and insertion rates of 0.0005 and 0.0001 for the low and high divergence case, respectively.

3 RESULTS

To evaluate the results of `HiFiX` and compare them to those of other methods, we used several benchmark sets of known homologous protein families. On each set of sequences, we first compared all proteins against all with BLASTP (Altschul *et al.*, 1997). The BLAST results were then analyzed with different clustering methods.

For each method, each reference family of the benchmark set was compared with the results of the clustering to identify the cluster with the highest number of corresponding sequences (target cluster). Each sequence of the reference family that was absent from the target cluster was counted as a FN. Each sequence of the target cluster that was present or absent from the reference family was counted as a true positive (TP) or a FP, respectively. We computed sensitivity TP/(TP+FN) and specificity TP/(TP+FP) for each cluster and report their values averaged over all reference families. We also report F-measure II, a weighted harmonic mean between sensitivity

Table 1. HiFiX performance compared with other clustering programs

Method	Nb. clusters	Spec.	Sens.	$\mathcal{F}$	Nb. clusters (≥10 proteins)	Spec.	Sens.	$\mathcal{F}$
	(a) on the Brown *et al.* (2006) benchmark of 866 enzymes Wittkop *et al.* (2010)				(b) on bacterial TCRRs (TCRR family)			
HiFiX	99	0.94	0.91	**0.90**	6	0.98	0.81	**0.88**
SiLiX	99	0.94	0.91	**0.90**	1	0.38	1.00	0.53
Louvain	109	0.95	0.79	0.82	32	1.00	0.57	0.71
MCL	47	0.84	0.98	**0.87**	18	0.68	0.92	0.77
TransClust	96	0.95	0.90	**0.91**	279	1.00	0.05	0.08
hcluster_sg	27	0.66	1.00	0.73	34	0.71	0.86	0.76

Spec: specificity; Sens: sensitivity; $\mathcal{F}$: F-measure II. $\mathcal{F}$ values >0.85 are in bold.

and specificity that was introduced in (Paccanaro *et al.*, 2006) [see details in Supplementary Material and also (Wittkop *et al.*, 2010)].

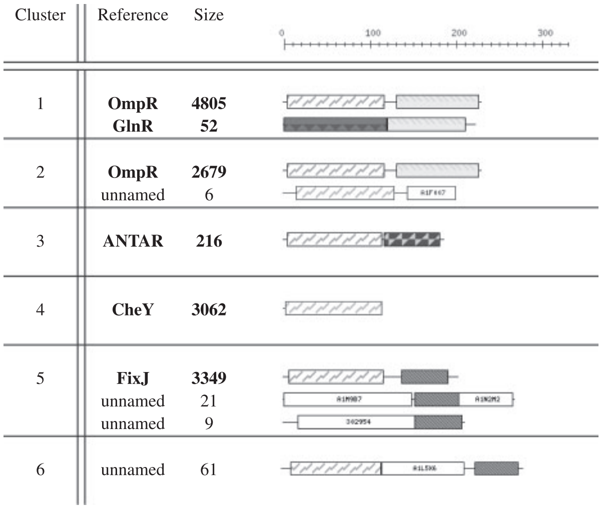
This procedure was used to evaluate SiLiX, used alone, or followed by the two additional HiFiX steps: step 2: decomposition of pre-families into communities with Louvain; step 3: hierarchical clustering into meta-communities that are the final families. To quantify the gain due to the third step, we also analyzed the results of the first two steps, i.e. SiLiX followed by Louvain, assuming that the found communities are the protein families.

For a comparison, we also performed the clustering with several other methods that were recently evaluated among the best currently available: MCL (Enright *et al.*, 2002), hcluster_sg (Ruan *et al.*, 2008) and TransClust (Wittkop *et al.*, 2010). We first tested these methods on sets of homologous protein families that have been manually analyzed by experts. Unfortunately, there are only a limited number of such benchmark sets and they tend to be biased toward relatively small families. Furthermore, even expert analyses cannot fully guarantee the absence of FP or FN in the reference set. To circumvent these limitations, we also performed tests on sets of simulated sequences for which, by construction, homology relationships are perfectly known.

3.1 Evaluation of clustering methods on known protein families

We first used a reference dataset published by (Brown *et al.*, 2006) containing 866 enzymes that were manually assigned to 91 protein families, and that was recently used to compare clustering methods by (Wittkop *et al.*, 2010). On this test set, all methods except hcluster_sg perform very well (Table 1a). The best results are obtained with TransClust, SiLiX and HiFiX. The results of HiFiX are identical to those of SiLiX because families generated by SiLiX are homogeneous on this dataset. It should be noted that these protein families are relatively small: the largest family contains only 215 sequences (average of 10 sequences per family). As discussed in the introduction, the risk of FPs due to alignment over-extension increases with the size of the family. Therefore this benchmark is a relatively easy case for clustering methods.

To evaluate the results of these methods on larger protein families, we analyzed a set of 14 260 bacterial TCRRs containing a conserved receiver domain and divergent output domains [see (Galperin, 2010) for a recent update]. This set of sequences (hereafter called TCRR) corresponds to the 4-th largest family in the HOGENOM database [(Penel *et al.*, 2009)] and has been very well studied. However its large size makes it more difficult to process correctly for clustering algorithms such as SiLiX and hence the resulting family is heterogeneous in HOGENOM (accession number HBG753323). Indeed an analysis of domain decomposition of its member proteins with ProDom (Bru *et al.*, 2005) indicates 10 distinct protein types (Fig. 3). HiFiX retrieves six families (Fig. 4) and performs much better than other methods (Table 1b). Compared with SiLiX, HiFiX is much more specific without losing much sensitivity. MCL performs quite well on this test set, but with a relatively high number of FPs compared with HiFiX. MCL gave many clusters, and notably clustered most single domain CheY-type proteins together with two-domain OmpR-type proteins, while other CheY-type proteins were scattered into a dozen clusters. hcluster_sg behaved similarly. The results of TransClust on this test set were poor in terms of sensitivity.



Cluster	Reference	Size
1	**OmpR** **GlnR**	**4805** **52**
2	**OmpR** unnamed	**2679** 6
3	**ANTAR**	**216**
4	**CheY**	**3062**
5	**FixJ** unnamed unnamed	**3349** 21 9
6	unnamed	61

Fig. 3. Correspondence between manually built clusters with domain arrangements from ProDom [(Bru *et al.*, 2005)] and HiFiX clusters on bacterial TCRRs (TCRR family).

3.2 Evaluation of clustering methods with simulated protein families

In order to assess the effects of sequence divergence and protein modularity on the behavior of HiFiX, we tested it further on a

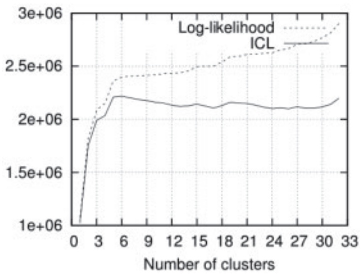


Fig. 4. ICL and $\log\mathcal{L}_Q(S,\widehat{Z};\widehat{\alpha},\widehat{\theta})$ on TCRR family, as a function of the number of clusters. Maximum ICL corresponds to the six clusters shown in Figure 3.

set of artificially evolved sequences. We simulated the evolution of protein sequences (250 amino-acid long), subject both to progressive divergence by accumulation of point mutations and to modular rearrangements. The progressive divergence was simulated using the tool `INDELible` (Fletcher and Yang, 2009), starting from a random ancestral sequence, and following a true phylogeny (based on 23S rRNA sequences) of 536 bacterial species (Fig. 5). We modulated the rate of sequence divergence by controlling the tree depth between 1.33 and 4.5 substitutions per site (which corresponds to an average of 32% identity between sequences from the outgroup and from the ingroup clades). Domain rearrangements were performed by the replacement of the C-terminal end of the protein by a non-homologous sequence: for a given clade of simulated sequences, i.e. deriving from a common ancestor, we exchanged the C-terminal end of each sequence with those of a different set of simulated sequences, species by species (Fig. 5). In our simulations, we introduced three independent events of domain replacements (i.e. we assumed that the rate of domain rearrangement was relatively low). We modulated the relative size of the rearranged C-terminal domain (x) from $x=0\%$ to $x=40\%$ of the protein length. Thus, when $x=0\%$, the whole set of simulated sequences corresponds to one single family of proteins, homologous over their entire length. When $x>0\%$, simulated sequences correspond to four different families: the three clades that derive from the nodes where the domain rearrangements occurred, and the rest of the sequences, that conserved the ancestral domain architecture. For each set of parameters (protein divergence rate, length of rearranged domain) we performed 20 simulations.

When protein divergence is low (average of 1.33 substitutions per site), both `TransClust` and `HiFiX` perform very well (Table 2). `hcluster_sg`, `MCL` and `SiLiX` give good results in the absence of domain rearrangement ($x=0\%$), but perform poorly when $x>0\%$ because they tend to cluster in the same family sequences with distinct domain architectures. In the case of `SiLiX`, this behavior is clearly due to over-extension of BLAST alignments beyond the homologous domain (e.g. see Figure 1), which leads to FP similarity links.

When protein divergence is high (average of 4.5 substitutions per site), `HiFiX` clearly outperforms other methods. The sensitivity of `Transclust` is very low with this set: as with the TCRR dataset analyzed above, it tends to create a large number of small homogeneous clusters. `MCL` works well for $x=0\%$ but is not robust to domain rearrangement. In conclusion `HiFiX` is the only method that remains robust both to sequence divergence and to domain rearrangements. We also performed simulations to test the impact of various factors on `HiFiX` clustering results (protein sequence length, unbalanced phylogenetic sampling, presence of proteins homologous to the shuffled domains). These analyses showed that `HiFiX` results are not affected by these factors (see Supplementary Material).

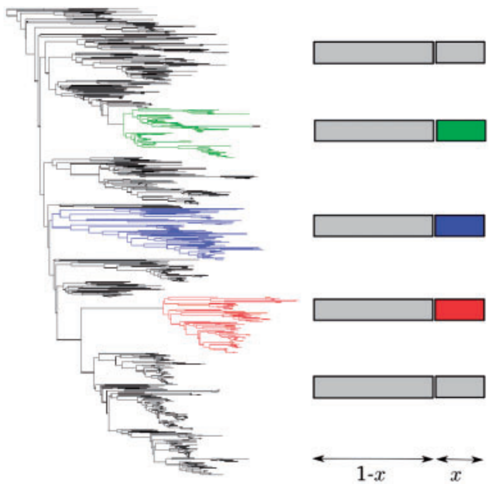


Fig. 5. The guide tree used in our simulations corresponds to a subset of a published phylogenetic tree of bacterial species, reconstructed from 23S rRNA (Pruesse *et al.*, 2007). Protein domain rearrangements were introduced in three nodes, by substituting the C-terminal fragment by a non-homologous sequence. These nodes correspond to the ancestors of the green, blue and red clades (containing, respectively, 165, 81 and 89 sequences). All other proteins (in black, 201 sequences) have conserved the ancestral domain architecture. We ran simulations with different values of x (the relative length of the C-terminal domain) and different rates of sequence evolution.

3.3 Impact of `SiLiX` parameters

Steps 2 and 3 of the `HiFiX` procedure do not require any parameter. However, the sensitivity of `HiFiX` depends on the parameters that are used by `SiLiX` in the first step to cluster sequences into pre-families. `SiLiX` requires two parameters: the minimal percentage of sequence identity s and the alignment coverage c, i.e. the fraction of protein length that is covered by the pairwise sequence alignment. We tested the impact of these parameters on `HiFiX` results (Table 3 and Supplementary Table S4). The `SiLiX` step is required to exclude from the similarity network (which will be analyzed by `Louvain` at the next step) all edges that only correspond to domain homology. We observed that if we do not set any constraint on the alignment coverage (c = 0%), then this leads to the clustering of sequences that are not homologous over their entire length (data not shown). However, `HiFiX` performances on our simulated dataset remain essentially unchanged for a wide range of relevant c coverage parameter (from $c=60\%$ to $c=90\%$, Supplementary Table S4). As expected, the use of stringent similarity parameter ($s\geq 40\%$) leads to a decrease in sensitivity. Indeed, it is necessary to use `SiLiX` criteria that are permissive enough to avoid the exclusion of relevant homology relationships at the first step. The use of more permissive `SiLiX` criteria requires more computing time (intrinsically due to the larger number of multiple alignments that have to be computed at step 3) but leads to a strong gain in sensitivity, with only limited loss in specificity for `HiFiX` . This indicates

Table 2. Average performance of clustering programs on simulated datasets of evolving modular proteins

Method	$x = 40\%$			$x = 25\%$			$x = 0\%$		
	Spec.	Sens.	$\mathcal{F}$	Spec.	Sens.	$\mathcal{F}$	Spec.	Sens.	$\mathcal{F}$
	Low divergence ($d = 1.33$)								
HiFiX	1.00	1.00	**1.00**	0.78	1.00	**0.88**	1.00	1.00	**1.00**
SiLiX	0.29	1.00	0.44	0.29	1.00	0.44	1.00	1.00	**1.00**
Louvain	1.00	0.98	**1.00**	0.91	1.00	**0.94**	1.00	0.98	**0.97**
MCL	0.29	1.00	0.44	0.29	1.00	0.44	1.00	1.00	**1.00**
TransClust	1.00	0.97	**0.98**	0.96	0.98	**0.96**	1.00	0.90	**0.94**
hcluster_sg	0.29	1.00	0.44	0.29	1.00	0.44	1.00	1.00	**1.00**
	High divergence ($d = 4.5$)								
HiFiX	1.00	0.95	**0.97**	0.93	0.94	**0.92**	1.00	0.91	**0.95**
SiLiX	0.95	0.95	**0.94**	0.62	0.95	0.71	1.00	0.99	**1.00**
Louvain	1.00	0.65	0.68	1.00	0.78	0.77	1.00	0.41	0.58
MCL	0.68	0.98	0.74	0.39	0.99	0.52	1.00	1.00	**1.00**
TransClust	1.00	0.29	0.43	1.00	0.28	0.42	1.00	0.10	0.19
hcluster_sg	0.41	1.00	0.55	0.31	1.00	0.46	1.00	1.00	**1.00**

Divergence is controlled by the guide tree depth d which is expressed in number of substitutions per site. x corresponds to the length fraction of heterologous domains (see text for details). Abbreviations are the same as in Table 1.

Table 3. Average performance of HiFiX on simulated datasets of modular proteins ($x = 40\%$) with SiLiX parameters set to varying percentages of sequence identity s and to a fixed alignment coverage $c = 80\%$

Method	$x = 40\%$		
	Spec.	Sens.	$\mathcal{F}$
	High divergence ($d = 4.5$)		
SiLiX $s = 25\%$	0.29	1.00	0.43
HiFiX	0.85	1.00	**0.90**
SiLiX $s = 30\%$	0.45	1.00	0.58
HiFiX	0.91	1.00	**0.94**
SiLiX $s = 35\%$ (default)	0.95	0.95	**0.94**
HiFiX	1.00	0.95	**0.97**
SiLiX $s = 40\%$	1.00	0.80	**0.88**
HiFiX	1.00	0.80	**0.88**
SiLiX $s = 50\%$	1.00	0.39	0.54
HiFiX	1.00	0.39	0.54

Details and abbreviations are the same as in Table 2.

that step 3 is efficient to reject FPs. Interestingly, HiFiX results remain stable for a relatively wide range of SiLiX parameters (in Table 3, from $s = 25\%$ to $s = 40\%$). Thus it appears that it is not necessary to use extremely permissive SiLiX criteria to obtain optimal results. Based on these simulations, we recommend using the default parameters of SiLiX ($s = 35\%$ and $c = 80\%$).

3.4 Computational footprint

While the main advantage of HiFiX is the quality of the clusters it generates, the associated computational expense is affordable even for very large datasets. Indeed the Louvain algorithm can handle huge networks with millions of vertices in reasonable time [see (Blondel *et al.*, 2008)]: finding communities can usually be achieved in only a few seconds. Second, mafft perfoms K multiple alignments on protein subsets and mafft-profile fewer than $l(K-1)$ alignments of alignments, so the overall alignment task scales linearly with K. Moreover mafft and mafft-profile present an excellent trade-off between speed and accuracy (Katoh *et al.*, 2009). Finally likelihood calculations rely on the very efficient HMMER3 package (Eddy, 2009). Altogether this makes HiFiX extremely efficient. Hence, using a single core Intel Xeon 3.07 GHz, we clustered the 536 simulated sequences in 5 s, the 866 enzymes in 12 s and the 14 260 TCRR proteins (a single pre-family with ≈6.4 million edges in the similarity network) in ∼75 min. As a benchmark, we analyzed a very large set of 3 206 033 sequences (extracted from the HOGENOM database version 6) displaying numerous pre-families with a power law distribution of size between 50 and 23 327 sequences. Taking advantage of multiprocessing to treat them in parallel on 24 cores Opteron 2.2 GHz, HiFiX required ∼13 700 min of calculation corresponding to <10 h elapsed time.

4 DISCUSSION

Proteins are the result of an evolutionary process, and can therefore naturally be classified into families of homologous sequences. We propose here a clustering method (HiFiX) that relies on three steps: (i) permissive clustering of sequences in pre-families; (ii) sub-clustering of pre-families into homogeneous clusters; and (iii) progressive merging of clusters into families, with evaluation of the quality of the multiple alignment at each step. The goal of this strategy is to maximize sensitivity (i.e. families should be as exhaustive as possible), while preserving the quality of the final multiple alignment (i.e. two clusters can be merged, as long as they are homologous over their entire length). The logics behind this strategy is to cluster all homologous sequences that are similar enough to obtain a reliable multiple alignment, from which it will be possible to construct a phylogenetic tree. We evaluated the quality of the clustering with several manually curated or simulated benchmark datasets. As expected, the clustering obtained after step 2 (Louvain) shows a higher specificity but lower sensitivity than after step 1 (SiLiX) (Tables 1 and 2). Interestingly, step 3 leads to recover a sensitivity almost as good as SiLiX, while preserving the specificity obtained at the second step with Louvain (Tables 1 and 2). The enzyme benchmark set from (Brown *et al.*, 2006) and the simulated set without modular evolution ($x=0$) correspond to relatively easy cases, where most tested methods, including SiLiX, give very good results. On these families, HiFiX performs as well as SiLiX, which indicates that steps 2 and 3 do not deteriorate the quality of the clustering. The more difficult cases correspond to families of multi-domain proteins that are not homologous over their entire length (the TCRR family and the simulated sets with $x>0$). HiFiX turned out to be the only method that is robust to such situations of modular evolution. We think that the main reason for these good results is that contrarily to other methods, that are all based on the analyses of pairwise sequence similarities, HiFiX uses information from the multiple alignment to decide whether a sequence should be included in a family or not. This step results in good specificity and therefore allows using more permissive parameters at the first step, which ensures good sensitivity. It should be noted that the results of the other tested methods might have been improved by tuning their parameters (here we used default parameters provided by the authors). However, as mentioned in the introduction, the optimal clustering parameters for a given set of families may in fact not be optimal for other families. Indeed, one important property of the HiFiX method is that it does not require setting a priori the same parameters for all families: the decision to include or not a sequence in a family is based on the analysis of the multiple alignment, and hence takes into account the pattern of evolution specific of that family. Of course, HiFiX sensitivity depends on the parameters used for step 1 (SiLiX). However, as long as SiLiX parameters are permissive enough, HiFiX results remain essentially unaffected.

HiFiX is designed to cluster sequences that are homologous over their entire length, by contrast to modular proteins that only share some homologous domains. HiFiX is therefore not appropriate to detect homology relationships among proteins that are only partially alignable. The choice of this relatively strict clustering criterion is primarily motivated by pragmatic considerations. Indeed, HiFiX was developed to be used for large scale phylogenomic studies, and notably for the construction of the HOGENOM database, which provides phylogenetic trees for thousands of gene families (Penel *et al.*, 2009). To be able to reconstruct the evolutionary relationships among homologous sequences, it is necessary to have a multiple alignment of good quality. Allowing the clustering of partially homologous proteins inevitably leads to low quality multiple alignments, which therefore produce erroneous phylogenetic trees. Of course, such cases could be corrected by manual expertise, but this cannot be done systematically for large scale phylogenomic analyses. An alternative approach would be to cluster homologous domains, instead of entire homologous proteins. The problem is that domains are generally short, and this strategy therefore leads to a loss of phylogenetic signal. The HiFiX clustering strategy thus reflects a tradeoff between the quantity and quality of phylogenetic information that can be recovered from large scale comparative sequence analyses. It should be noted however that despite this intrinsic limitation, HiFiX performs at least as well as other methods on manually curated benchmark sets (Table 1). This is most probably because these methods, although they do not set a priori constraints on the alignment length, have to use relatively strict parameters to limit the rate of FPs.

In conclusion, HiFiX provides a significant improvement in clustering quality over other existing methods, notably on multi-domain protein families, which represent a large fraction of all sequences. The HiFiX procedure is fast enough to be used in practice on very large sequence data sets.

ACKNOWLEDGEMENTS

The authors would like to give special thanks to David Robelin for his inspiration, and would like to thank Bruno Spataro and colleagues for the computing facilities, and also Thomas Bernard, Catherine Matias, Bernard Prum and lastly Tobias Wittkop for answers about the use of his program. We also thank three anonymous reviewers for their constructive comments.

Funding: French Agence Nationale de la Recherche under grants NeMo ANR-08-BLAN-0304-01 and ANCESTROME ANR-10-BINF-01-01; and European Union under grant IMPACT FP7-RI-213037.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrade,R.F. *et al.* (2011) Detecting network communities: an application to phylogenetic analysis. *PLoS Comput. Biol.*, **7**, e1001131.

Apeltsin,L. *et al.* (2011) Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution. *Bioinformatics*, **27**, 326–333.

Atkinson,H.J. *et al.* (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE*, **4**, e4345.

Biernacki,C. *et al.* (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 719–725.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.-Theory E.*, **2008**, P10008+.

Brown,S.D. *et al.* (2006) A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.*, **7**, R8.

Bru,C. *et al.* (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **33**, D212–D215.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eddy,S. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Fletcher,W. and Yang,Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, **26**, 1879–1888.

Fokkens,L. *et al.* (2010) Enrichment of homologs in insignificant BLAST hits by co-complex network alignment. *BMC Bioinformatics*, **11**, 86.

Fortunato,S. (2010) Community detection in graphs. *Phys. Rep.*, **486**, 75–174.

Galperin,M.Y. (2010) Diversity of structure and function of response regulator output domains. *Curr. Opin. Microbiol.*, **13**, 150–159.

Girvan,M. and Newman,M.E. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.

Gonzalez,M.W. and Pearson,W.R. (2010) Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.

Han,K.J. *et al.* (2008) Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. *IEEE T Audio Speech*, **16**, 1590–1601.

Katoh,K. *et al.* (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.*, **537**, 39–64.

Medini,D. *et al.* (2006) Protein homology network families reveal step-wise diversification of Type III and Type IV secretion systems. *PLoS Comput. Biol.*, **2**, e173.

Miele,V. *et al*. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.

Nowicki,K. and Snijders,T.A.B. (2001) Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, **96**, 1077–1087.

Paccanaro,A. *et al*. (2006) Spectral clustering of protein sequences. *Nucleic Acids Res.*, **34**, 1571–1580.

Penel,S. *et al*. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, **10** (Suppl. 6), S3.

Picard,F. *et al*. (2009) Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics*, **10** (Suppl. 6), S17.

Pruesse,E. *et al*. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.

Ruan,J. *et al*. (2008) TreeFam: 2008 update. *Nucleic Acids Res.*, **36**, D735–D740.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Song,N. *et al*. (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.*, **4**, e1000063.

Tatusov,R.L. *et al*. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.

Vilella,A.J. *et al*. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.

Wittkop,T. *et al*. (2010) Partitioning biological data with transitivity clustering. *Nat. Methods*, **7**, 419–420.

Zhang,S.B. *et al*. (2011) Phylogeny inference based on spectral graph clustering. *J. Comput. Biol.*, **18**, 627–637.