Bioinformatics Advance Access published August 11, 2016

Bioinformatics, 2016, 1–9
doi: 10.1093/bioinformatics/btw213
Advance Access Publication Date: 22 April 2016
Original Paper

Sequence analysis

Bayesian nonparametrics in protein remote homology search

Mindaugas Margelevičius

Institute of Biotechnology, Vilnius University, Vilnius 10257, Lithuania

Associate Editor: Burkhard Rost

Received on November 11, 2015; revised on March 8, 2016; accepted on April 14, 2016

Abstract

Motivation: Wide application of modeling of three-dimensional protein structures in biomedical research motivates developing protein sequence alignment computer tools featuring high alignment accuracy and sensitivity to remotely homologous proteins. In this paper, we aim at improving the quality of alignments between sequence profiles, encoded multiple sequence alignments. Modeling profile contexts, fixed-length profile fragments, is engaged to achieve this goal.

Results: We develop a hierarchical Dirichlet process mixture model to describe the distribution of profile contexts, which is able to capture dependencies between amino acids in each context position. The model represents an attempt at modeling profile fragments at several hierarchical levels, within the profile and among profiles. Even modeling unit-length contexts leads to greater improvements than processing 13-length contexts previously. We develop a new profile comparison method, called COMER, integrating the model. A benchmark with three other profile-to-profile comparison methods shows an increase in both sensitivity and alignment quality.

Availability and Implementation: COMER is open-source software licensed under the GNU GPLv3, available at https://sourceforge.net/projects/comer.

Contact: mindaugas.margelevicius@bti.vu.lt

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

1 Introduction

Proteins, large molecules, perform a huge number of functions within living organisms. Given that the three-dimensional (3D) structure of a protein determines its function, Structural Genomics centers aim and dominant programs such as the Protein Structure Initiative (Montelione, 2012) aimed at the determination of the 3D structure for most proteins in nature to facilitate understanding of how living organisms function. With increasing coverage of protein fold space (Khafizov *et al.*, 2014), these efforts are greatly assisted by building protein models by homology to solved 3D structures (Schwede, 2013). Homology discovery founded on data of solved 3D structures, therefore, constitutes the foundation for homology modeling. Results of the Critical Assessment of Structure Prediction (CASP) experiments (Moult *et al.*, 2014) reveal homology search methods as some of the most decisive intermediate means in protein 3D structure prediction. Based on alignments of sequence profiles (generalizations of protein sequence families), these methods (Söding and Remmert, 2011) can reliably establish homologous relationships even between proteins sharing nearly random (6–8%) sequence identity. Nevertheless, such relations are remotely homologous and are often missed. Increasing the sensitivity to remotely homologous proteins and the accuracy of alignments is accordingly of great practical importance.

To improve the quality of the sequence profile and to obtain the probability distribution of amino acids (residues) more relevant to the sequence family, many efforts were put into the calculation of mutation (substitution) probabilities, including the development of the pseudo-counts (Altschul *et al.*, 2009; Sjölander *et al.*, 1996), composition-based statistics (Schäffer *et al.*, 2001) and compositional adjustment (Yu and Altschul, 2005) and mixing with many other sequence/structure-derived measures (see, e.g. Yan *et al.*, 2013). The advances in the calculation of substitution probabilities at each profile position have induced more accurate profile alignments and consequently, increased sensitivity to remote homologs

(Kryshtafovych *et al.*, 2014). Despite the apparent progress in developing profile-to-profile comparison methods, some problems remain. Alignment accuracy, one of the most important challenges in homology modeling (Venclovas and Margelevičius, 2009), seems to have saturated recently (Kryshtafovych *et al.*, 2014) and further improvements in alignment necessitate.

Enhancing the description of substitution probabilities so that the probabilities accurately reflect a structure/sequence environment (context) addresses the problem. Biegert and Söding (2009) showed that generalizing substitution probabilities for each position by mixing them with a library of context profiles has increased the sensitivity of the profile-to-sequence alignment method PSI-BLAST (Altschul *et al.*, 1997) by more than 64%. The context profiles—representative profile fragments—were not derivatives of structure information but derived from clustering profile fragments. The rationale of using sequence-based contexts was that the profile fragments with different distributions of substitution probabilities could adopt similar local structures. If grouped together, these profile fragments would lose their specificity.

Biegert and Söding (2009) modeled the profile fragment by using the product of weighted multinomial probabilities and optimized the substitution probabilities of each context profile with the Expectation Maximization algorithm. In that framework, the number of context profiles is a parameter whose value was found in a homology detection benchmark procedure.

In this work, we aim at increasing the sensitivity of profile comparisons and the accuracy of alignments, addressing the issue of profile contexts (Supplementary Fig. S1) from a Bayesian non-parametrics perspective. We propose a hierarchical Dirichlet process mixture model to describe the distribution of profile contexts and apply the inferred mixtures to the problem of protein homology detection. The hierarchical Dirichlet process (HDP) is developed (Teh *et al.*, 2006) for model-based clustering of grouped data, where each group is associated with a mixture model. The essential property of the HDP is that it employs a Dirichlet process (DP) for every group of data, where all the DPs share a base distribution which itself is drawn from a Dirichlet process. Such a construction enables sharing of clusters across groups and expresses the property of profiles (sequence families) of sharing common features. Considering profiles as separate groups, the HDP mixture model formalizes the purpose of finding characteristic clusters of contexts within each profile with sharing of the clusters among them. An important feature of the HDP mixture model is that the number of clusters is unknown and is to be inferred from the data in Markov chain Monte Carlo (MCMC) sampling. Overgeneralization of profile contexts leads to signal loss and reduced specificity, and this feature allows to avoid of selecting a particular number of clusters or placing an explicit parametric prior on it.

To our knowledge, HDP-based non-parametric Bayesian models to define the distribution of profile contexts or substitution probabilities have not been considered in the literature. Nguyen *et al.* (2013), though, modeled multiple sequence alignment (MSA) columns with DP Dirichlet mixtures. They specified residue frequencies at single columns by a Dirichlet distribution and inferred the Dirichlet mixtures from MSA data. Although computationally convenient, such a specification has some limitations. One is that a multivariate random variable drawn from a Dirichlet distribution corresponds to a scaled vector of independently distributed *gamma* random variables. The Dirichlet distribution, therefore, lacks versatility to capture inter-component correlations as illustrated in Figure S2 of the Supplementary Materials.

We build the model to handle profile data. The profile generalizes each MSA column with a vector of substitution probabilities, which depending on sequence weighting scheme, embody soft dependencies between adjacent columns. With these properties, substitution probabilities in our setting constitute an observation vector. To control correlations between residues, we model the distribution of substitution probabilities with a logistic-normal distribution (Aitchison and Shen, 1980), which follows from the logistic transformation applied to a normal distribution and, to the best of our knowledge, publicly has not been considered to characterize probabilities in profiles. The context (Supplementary Fig. S1) comprises an array of such vectors and is a random variable whose distribution the proposed mixture model describes. The model is valid for contexts of any length. Although, in this work, we confine to the context of one position, Section 5 discusses how the model without any corrections adapts to longer contexts.

2 Methods

2.1 Bayesian mixture model

This section defines the distribution of profile contexts. Let us consider, without loss of generality (see Section 5 for a discussion of the general case), a context of length one, i.e. consisting of a single vector of substitution probabilities. The positive real-valued substitution probabilities for all 20 residues sum to 1, but only $A = 19$ of them are independent. We assume the A-dimensional vector $\boldsymbol{f} = (f_1, \ldots, f_A)^{\mathrm{T}}$, $f_{A+1} = 1 - \sum_{a=1}^{A} f_a$, of substitution probabilities to be distributed according to a logistic-normal distribution, which is defined on the strictly positive simplex (Aitchison and Shen, 1980). The requirement of strictly positive probability values is always met as profile substitution probabilities result from mixing observed residue frequencies with pseudo-counts (Altschul *et al.*, 2009). The inverse logistic transformation applied to $\boldsymbol{f}$, $log(\boldsymbol{f}/f_{A+1})$, produces a normally distributed variable $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_A)^{\mathrm{T}}$. We thus focus on the distribution of the normal variables.

Let i indexes each context within a particular profile j, $j = 1, \ldots, R$. Then, profile contexts $\boldsymbol{\eta}_{ji} = (\eta_{ji1}, \ldots, \eta_{jiA})^{\mathrm{T}}$ are assumed to arise as draws from a normal distribution. We place prior distributions on the parameters of the mean vector $\boldsymbol{\mu}_{ji}$ and covariance matrix $\boldsymbol{\Sigma}_{ji}$ of the normal variable $\boldsymbol{\eta}_{ji}$. The prior distribution for the covariance matrix $\boldsymbol{\Sigma}_{ji}$ is specified as an inverted Wishart distribution with $\nu_0 + A + 1$ $(\nu_0 > A - 1)$ degrees of freedom and scale matrix $\boldsymbol{\Lambda}_0$, and for the mean vector $\boldsymbol{\mu}_{ji}$ as a normal distribution with mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_{ji}/\kappa_0$ scaled by real $\kappa_0 > 0$:

$$\begin{aligned} \boldsymbol{\Sigma}_{ji} | \nu_0, \boldsymbol{\Lambda}_0 &\sim \mathcal{IW}_A(\nu_0 + A + 1, \boldsymbol{\Lambda}_0), \\ \boldsymbol{\mu}_{ji} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{ji}, \kappa_0 &\sim \mathcal{N}_A(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{ji}/\kappa_0). \end{aligned} \tag{2.1}$$

The parameter vector $\boldsymbol{\theta}_{ji} = (\boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji})$ of each profile context $\boldsymbol{\eta}_{ji}$ can be regarded as being itself a random variable, and their values, in particular, are not necessarily distinct.

Commonly, profiles share contexts (local features), including even those representing proteins or their families from different classes. Moreover, similar contexts can appear multiple times within the same profile, and the extent of similar contexts depends on and varies across profiles. To include these properties in our model, we first introduce a prior distribution G_j for the parameters $(\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2}, \ldots)$ of the corresponding context observations $(\boldsymbol{\eta}_{j1}, \boldsymbol{\eta}_{j2}, \ldots)$, unique to each profile j. We assume that the parameters are conditionally independent given G_j, and G_j itself is a random probability measure distributed as a Dirichlet process (Ferguson, 1973), $\mathrm{DP}(\tau_0, G_0)$, with concentration parameter τ_0 and base probability measure G_0 for the parameters $(\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2}, \ldots)$. The discrete nature of the G_j (Ferguson, 1973) allows context sharing within profile j by associating context observations with the same values of the parameters $\boldsymbol{\theta}_{ji}$. To permit sharing across profiles, the G_0 is modeled as a global random probability measure also distributed according to a DP with concentration parameter γ and base probability measure F. Now, the base measure F provides the prior distribution for the parameters $\boldsymbol{\theta}_{ji}$ across all profiles.

According to this setting, known as the hierarchical Dirichlet process (Teh *et al.*, 2006), the distribution G_0 varies around the prior F and the G_j around the G_0, where the hyperparameters γ and τ_0, respectively, control the amount of variability at the two hierarchical levels. Exactly this sharing of the parameter atoms (values) by both the global probability measure G_0 and the distribution G_j assigned its own set of weights for those parameters at the profile level (Supplementary Fig. S3), that makes the HDP model different from the corresponding DP model that provides all profiles with a single global set of weights. We show in the simulation study in Section 2.3 that the HDP model significantly outperforms its DP counterpart.

Under the assumptions above, we have the following HDP mixture model:

$$\begin{aligned} G_0|\gamma, F &\sim \mathrm{DP}(\gamma, F), \quad & G_j|\tau_0, G_0 &\sim \mathrm{DP}(\tau_0, G_0), \\ \boldsymbol{\theta}_{ji}|G_j &\sim G_j, & \boldsymbol{\eta}_{ji}|\boldsymbol{\theta}_{ji} &\sim \mathcal{N}_A(\boldsymbol{\theta}_{ji}). \end{aligned} \tag{2.2}$$

As follows from the connection of the DP with a Pólya urn scheme (Blackwell and MacQueen, 1973), the probability for $\boldsymbol{\theta}_{ji}$ to take an already observed value, given all the other random variables $\{\boldsymbol{\theta}_{ji'}\}_{i'\neq i}$ and the base measure F, where the G_j and G_0 have been integrated out, is proportional to the size of the cluster containing variables with the given value. We have the hierarchical arrangement of the DPs, and the clustering property exhibits at both levels.

Consider, first, the top-level DP in (2.2). Let $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ denote iid random variables corresponding to K distinct values of the parameters $\{\boldsymbol{\theta}_{ji}\}_{j,i}$, distributed according to F. Let also $\boldsymbol{\psi}_{j1}, \ldots, \boldsymbol{\psi}_{j,m_{j\cdot}}$ over all j be a Pólya sequence (Blackwell and MacQueen, 1973) of random variables with parameter γ and probability measure F, where $m_{j\cdot}$ is the number of these variables in profile j. Then, the conditional distribution of $\boldsymbol{\psi}_{jl}$ given $\{\boldsymbol{\psi}_{jl'}\}_{j,l'\neq l}$ and F is

$$\boldsymbol{\psi}_{jl}|\{\boldsymbol{\psi}_{jl'}\}_{j,l'\neq l}, \gamma, F \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma}\delta_{\boldsymbol{\phi}_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma}F, \tag{2.3}$$

where $m_{\cdot k}$ is the number of variables $\boldsymbol{\psi}_{jl'}$ that are equal to $\boldsymbol{\phi}_k$ over all j and $l' \neq l$, $m_{\cdot\cdot}$ (dots represent marginal counts) is the total number of variables over all j and $l' \neq l$, and δ_x denotes the unit measure concentrating at x. The distribution (2.3) corresponds to the distribution $\boldsymbol{\psi}_{jl} \sim G_0$, $G_0 \sim \mathrm{DP}(\gamma, F)$, with G_0 integrated out. Each variable $\boldsymbol{\psi}_{jl}$ takes on the value $\boldsymbol{\phi}_k$ with probability proportional to $m_{\cdot k}$ and takes a new value from F with probability proportional to γ (we refer the interested reader to Teh *et al.*, 2006, for a detailed description of the HDP).

At the lower level of the DPs, in each profile j individually, we obtain the conditional distribution of the random variable $\boldsymbol{\theta}_{ji}$ given $\{\boldsymbol{\theta}_{ji'}\}_{i'\neq i}$ and G_0, when G_j has been integrated out, concentrated on $\{\boldsymbol{\psi}_{jl}\}_l$

$$\boldsymbol{\theta}_{ji}|\{\boldsymbol{\theta}_{ji'}\}_{i'\neq i}, \tau_0, G_0 \sim \sum_{l=1}^{m_{j\cdot}} \frac{n_{jl\cdot}}{n_{j\cdot\cdot} + \tau_0}\delta_{\boldsymbol{\psi}_{jl}} + \frac{\tau_0}{n_{j\cdot\cdot} + \tau_0}G_0 \tag{2.4}$$

for all j, where $n_{jl\cdot}$ is the number of variables $\boldsymbol{\theta}_{ji'}$ equal to $\boldsymbol{\psi}_{jl}$ over all $i' \neq i$ in profile j, and $n_{j\cdot\cdot}$ is the total number of variables over all $i' \neq i$ in profile j. For each j, the variables $\boldsymbol{\theta}_{ji}$ take on values $\boldsymbol{\psi}_{jl}$ with probability proportional to $n_{jl\cdot}$ and take a new value from G_0 with probability proportional to τ_0. Since G_0 is distributed as the top-level DP, integrating it out gives the conditional distribution of $\boldsymbol{\theta}_{ji}$ expressed in terms of the variables $\boldsymbol{\phi}_k$ and the base probability measure F

$$\begin{aligned} \boldsymbol{\theta}_{ji}|\{\boldsymbol{\theta}_{ji'}\}_{i'\neq i}, \tau_0, \gamma, F \sim &\sum_{k=1}^{K} \frac{n_{j\cdot k}(m_{\cdot\cdot} + \gamma) + \tau_0 m_{\cdot k}}{(n_{j\cdot\cdot} + \tau_0)(m_{\cdot\cdot} + \gamma)}\delta_{\boldsymbol{\phi}_k} \\ &+ \frac{\tau_0\gamma}{(n_{j\cdot\cdot} + \tau_0)(m_{\cdot\cdot} + \gamma)}F \quad \text{for all } j, \end{aligned} \tag{2.5}$$

where $n_{j\cdot k}$ is the number of variables $\boldsymbol{\theta}_{ji'}$ equal to $\boldsymbol{\phi}_k$ over all $i' \neq i$ in profile j. Note that m_{jk} is the counter associated with the variables $\boldsymbol{\psi}_{jl}$ while n_{jlk} with the $\boldsymbol{\theta}_{ji}$. Thus, m_{jk} counts the number of *local* clusters associated with value $\boldsymbol{\phi}_k$ in profile j while $n_{j\cdot k}$ records the overall number of $\boldsymbol{\theta}_{ji}$ in these clusters in profile j. Local clusters are connected across all profiles through sharing the atoms $\boldsymbol{\phi}_k$ that determine *global* clusters. In other words, according to the model (2.2), profile context observations $\boldsymbol{\eta}_{ji}$ associated with parameters $\boldsymbol{\theta}_{ji}$ group into clusters within each profile j, which are mutually connected among all profiles. This scheme of the HDP is equivalent to the Chinese restaurant franchise metaphor presented by Teh *et al.* (2006).

Note that the iid variables $\{\boldsymbol{\theta}_{ji}\}_i$ are exchangeable and hence (2.4) and (2.5) are applicable for all i. We exploit this property in posterior inference.

2.2 Inference

Let us introduce the indicator variables $\{\ell_{ji}\}_{j,i}$ and $\{\varkappa_{jl}\}_{j,l}$. The variable ℓ_{ji} indicates local cluster l in profile j, with which the observation vector $\boldsymbol{\eta}_{ji}$ is associated: $\ell_{ji} = (j, l)$ implies $\boldsymbol{\theta}_{ji} = \boldsymbol{\psi}_{jl}$. Whereas $\varkappa_{jl}$ provides the index k of the global cluster involving local cluster l in profile j: $\varkappa_{jl} = k$ implies $\boldsymbol{\psi}_{jl} = \boldsymbol{\phi}_k$. Exploiting that the ℓ_{ji} and $\varkappa_{jl}$ exhibit the same exchangeability properties as the corresponding variables $\boldsymbol{\theta}_{ji}$ and $\boldsymbol{\psi}_{jl}$, our MCMC sampling scheme deals with the indicator variables. A single iteration of the posterior sampling algorithm consists of the following steps: 1. Update ℓ_{ji} for all j and i. 2. Update $\varkappa_{jl}$ for all j and l. 3. Perform Metropolis updates repeatedly for a number of times. 4. Update the hyperparameters γ, τ_0, κ_0 and ν_0. Section S1 of the Supplementary Materials describes each step in detail, introduces a split-merge algorithm implementing the Metropolis updates in step 3, and presents a parallel MCMC sampling algorithm.

2.3 Simulation study

Section S2 of the Supplementary Materials investigates the performance of the MCMC sampling procedure defined in Section 2.2. We constructed $K = 200$ distinct clusters each containing $n_{\cdot\cdot k} = 200$ observations generated from a logistic-normal distribution. We grouped all observations into $R = 200$ groups (profiles), initiated three MCMC sequences initialized with three different starting configurations, and performed convergence diagnostics. The MCMC sequences converged. We obtained a strong correlation between the estimated and true values of the parameters used to generate the data. Section S2 also demonstrates (Supplementary Fig. S5) the advantage of using the HDP mixture model over the corresponding DP mixture model for our experiments. We refer the reader to Section S2 for a detailed discussion of the experiments.

3 Model application

The application of the mixture model (Section 2.1) to protein homology detection consists of several parts. First, we compose a large collection of non-redundant data of profile contexts. Next, we infer the model parameters from the data. Then we apply the model to derive scores to be incorporated into a comparison of profiles. Lastly, we evaluate the effect of the model-derived scores on the sensitivity and alignment quality of the profile-to-profile comparisons. To accomplish all these steps, we developed profile construction and comparison applications which combined the main features of the methods COMA (Margelevičius and Venclovas, 2010) and HHsearch (Söding, 2005). The features include the profile construction protocol and compositional adjustments of the first method, and transition probabilities in the context of hidden Markov model (HMM) and log-sum-of-odds scoring of the second one. We call the new method COMER (profile COMparER).

The first two steps, that is how the profile context data were compiled and the inference process, are described in Section S3.1. We obtained a final set of $R = 49\,332$ profiles, where R also corresponds to the number of groups in the HDP mixture model. For each profile $j = 1, \ldots, R$, a number of contexts, proportional to the profile length, were randomly sampled, contributing to a total of $n_{\ldots} = 991\,833$ distinct observations. We ran the sampler on the data and found a maximum likelihood estimate of the state, $\boldsymbol{c}^{\mathrm{opt}}$, with the number of global clusters $K = 1260$ and the total number of local clusters $m_{\cdot\cdot} = 962\,446$. The distributions of observations in most populated clusters, as expected, approach multivariate normal distributions. Supplementary Figure S7 represents the most populated cluster.

We emphasize that clusters are characterized by the parameters of the t-distribution, calculated purely from data (Appendix A.1). No fit of parameters to the data takes place. Note also that the numbers of clusters were inferred from the data. We employ the mixture model corresponding to the state $\boldsymbol{c}^{\mathrm{opt}}$ for deriving scores. (State samples around the maximum likelihood state $\boldsymbol{c}^{\mathrm{opt}}$ yielded similar results [Section 4], meaning that $\boldsymbol{c}^{\mathrm{opt}}$ is stable with the majority of clusters being of high likelihood [Supplementary Fig. S6].)

3.1 Derivation of scores

Since our evaluation system (see Section 4) is based on the assessment of the quality of generated protein 3D structural models, we aim at improving the accuracy of profile alignments by deriving the scores from structural information as follows. For each sequence from a subset of protein domains with known 3D structures from the SCOP database (Murzin *et al.*, 1995) (version 1.71) filtered to 20% sequence identity, we calculated profiles from MSAs obtained by running 2 iterations of HMMER3 (Eddy, 2011) against the UniRef50 (The UniProt Consortium, 2014) sequence database. For every position of each of 4611 built profiles, the posterior distribution of cluster membership for the context $\boldsymbol{\eta}^{\star}$ characterizing a single profile position was computed in this way: The posterior probability for $\boldsymbol{\eta}^{\star}$ to belong to global cluster k is $p(\varkappa^{\star} = k | \boldsymbol{\eta}^{\star}, \boldsymbol{c}^{\mathrm{opt}}, \mathbf{H}) \propto p(\varkappa^{\star} = k | \boldsymbol{c}^{\mathrm{opt}}) p(\boldsymbol{\eta}^{\star} | \mathbf{H}_k)$, where $\varkappa^{\star}$ is an indicator variable for $\boldsymbol{\eta}^{\star}$, $\mathbf{H}$ denotes the complete set of observations, $\mathbf{H}_k$ is a matrix of the observations associated with cluster k, $p(\boldsymbol{\eta}^{\star} | \mathbf{H}_k)$ represents a t-distribution (see Section S1.1 and Appendix A.2 of the Supplementary Materials), and $p(\varkappa^{\star} = k | \boldsymbol{c}^{\mathrm{opt}})$ is the prior probability of mixture component k. For $n_{j\cdot\cdot}$ does not vary greatly across all profiles j, we approximate $n_{j\cdot\cdot} = n_{\ldots}/R$ and calculate $p(\varkappa^{\star} = k | \boldsymbol{c}^{\mathrm{opt}})$ by averaging the prior probability of mixture component k over all j. Then, using (2.5), we get (see Section S3.2 for the derivation)

$$p(\varkappa^{\star} = v | \boldsymbol{\eta}^{\star}, \boldsymbol{c}^{\mathrm{opt}}, \mathbf{H}) \propto \begin{cases} [n_{\cdot\cdot k}(m_{\cdot\cdot} + \gamma) + R\tau_0 m_{\cdot k}] p(\boldsymbol{\eta}^{\star} | \mathbf{H}_k) & \text{if } v = k, \\ R\tau_0 \gamma p(\boldsymbol{\eta}^{\star}) & \text{if } v = k^{\star}, \end{cases} \tag{3.1}$$

where $k^{\star}$ denotes an unrepresented (new) global cluster, and $p(\boldsymbol{\eta}^{\star})$ is the prior probability of $\boldsymbol{\eta}^{\star}$ distributed as the t-distribution (S1.7) defined in Section S1.2 of the Supplementary Materials. Note that while (3.1) predicts global cluster membership, the structure of local clusters is taken into account by the variables m_{jk}.

Next, the 4611 protein domains were compared all against all with DALI (Holm *et al.*, 2008), a method for pairwise comparison of protein structures. Using only statistically significant matches (DALI Z-score ≥ 3) of the pairwise structural alignments (duplicate pairs were removed), we derived MSAs for each domain and weighted (Henikoff and Henikoff, 1994) sequences within each MSA to reduce redundancy. Having computed by (3.1) the distribution predicting cluster membership for a context for each position of the sequences in the MSAs, we calculated weighted frequencies of cluster pairs that had arisen from the aligned sequences in the MSAs. Each sequence position, though, considered a single cluster determined by $\arg\max_v p(\varkappa^{\star} = v | \boldsymbol{\eta}^{\star}, \boldsymbol{c}^{\mathrm{opt}}, \mathbf{H})$. In addition, to ensure a reliable subset of the data, we analyzed the aligned positions which had paired residues in a structural superposition and predicted clusters with probability at least 0.5. The score of having aligned clusters x and y was then calculated as $\log[(f_{xy} + f_{yx})/(2g_x g_y)]$, where f_{xy} is the weighted frequency of occurrence for pair x, y and $g_x = \sum_y f_{xy}$. We thus obtained a $K \times K$ score table (see Section S3.3 for more details and discussion of the validation of the scores).

The comparison of a pair of profiles employs the score table for each pair of positions of two profiles, considered for producing an alignment, where a pair of cluster indices defined as described above selects an entry from the score table. The selected score is down-weighted and is added to the total score used to score a pair of profile positions. We emphasize that while the posterior inference process is computationally intensive, the application of the scores is fast and does not take significant time in profile comparison.

4 Results

To evaluate the efficiency of the derived scores, we performed an extensive comparison of sequence profiles of protein domains. Protein domains served as an evaluation basis for two reasons. First, available 3D domain structures enable to assess reliably the accuracy of profile alignments, a major determinant in protein modeling. Second, domain sequences used to produce profiles reduce the probability of the homologous over-extension error (Gonzalez and Pearson, 2010) leading to a deterioration of the quality of profiles. Here, the dataset comprised the protein domains from the SCOPe database (Fox *et al.*, 2013) (version 2.03) filtered to 20% sequence identity to challenge remote homology detection. The resulting domains excluding the members of class g of small proteins were divided evenly into training and test set of 4916 and 4915 representatives, respectively. Training set was used to optimize the application parameters (weights) for the score table, while the test set to assess the performance on previously unseen data. Multiple sequence alignments for the construction of profiles for domain sequences from both sets resulted from searching each sequence with PSI-BLAST (Altschul *et al.*, 1997) (version 2.2.28+) using a sequence inclusion threshold of 10^{-5} and soft masking of low complexity regions for 6 iterations against the UniRef50 sequence database (downloaded 10/24/2013).

Our benchmark tests included three additional profile-to-profile comparison methods: HHsearch (Söding, 2005) (version 2.0.16, released 2013), PPAS (Yan *et al.*, 2013) from the I-TASSER software package (Roy *et al.*, 2010) and FFAS (Jaroszewski *et al.*, 2011). (We wanted to include more state-of-the-art methods into the benchmark, but most of them are designed as internet servers or do not distribute computer utilities to make profiles from user-supplied MSAs.) All the methods except FFAS exploit secondary structure (SS) predictions. The new method COMER and HHsearch incorporated SS predictions obtained by PSIPRED (Jones, 1999). Whereas SS predictions calculated by PSSpred from the I-TASSER package accompanied the PPAS sequence profiles. For each method, we constructed profiles using the same set of MSAs, compared the profiles all against all, and evaluated the sensitivity of detecting related protein pairs and the accuracy of the produced alignments. We found that HHsearch performed better with the option of posterior probability threshold for maximum accuracy alignment set to 0.3 (-mact 0.3). The evaluation thus used this option value for HHsearch. In addition, as HHsearch employs 13 −length contexts in comparison of profiles, we evaluated the effect of calculating contexts by both methods COMER and HHsearch.

Each method was evaluated by considering top-ranked profile alignments sorted by reported their statistical significance. Only one of two alignments between the same pair of different domains, that with higher statistical significance, was retained. Alignments between identical domains were removed.

4.1 Sensitivity

Although the SCOPe database used here for the dataset of protein domains classifies domains into the families, superfamilies, folds and classes, we do not trust the classification blindly in our evaluation. Protein domains in SCOP(e) are classified according to the criteria of common evolutionary origin and structural similarity. However, the protein structure space, as a number of studies have shown (see, e.g. Sadreyev *et al.*, 2009), is continuous, meaning that one could expect to discover many relationships between proteins, conflicting with those established in the classification. We thus consider a pair of aligned domains as a correct match (true positive, TP) if the domains belong to the same superfamily *or* their structural similarity by DALI is statistically significant (Z-score ≥ 2). Otherwise, the relationship implied by an aligned pair is considered spurious (false positive, FP).

Figure 1 shows the number of true positives versus the number of false positives for each evaluated method. The new method COMER at a false discovery rate of 10% detects 103% more related pairs than PPAS, 16% than FFAS and 8% than HHsearch. By a permutation test for unpaired data (Venkatraman, 2000), the COMER curve in the figure differs significantly from the curves of the other methods at the 1% level. The scores derived from describing unit-length contexts by the model (2.2) increased the sensitivity of COMER for the top 60 000 alignments by 5.3%. In contrast, for the same amount of top alignments, HHsearch advanced by 2.5% by computing 13−length contexts.

Additionally, to prevent from any possible misinterpretations related to classifying pairwise alignments as false positives, we considered each aligned pair of structurally dissimilar (DALI Z-score < 2) domains from different superfamilies but from the same fold as being of unknown relationship. Ignoring these 'unknown' relationships ensures that possibly related pairs from the same fold, which by definition groups structurally similar superfamilies, are not wrongly classified as errors. The revised results, shown in Figure S10 of the Supplementary Materials, display no major change in sensitivity for the methods, meaning that among the top profile alignments, there are few aligned pairs belonging to different superfamilies but the same fold and for which DALI Z-score < 2.

A more stringent criterion for a true positive, formulated as the requirement for a pair to belong to the same superfamily or to have a DALI Z-score ≥ 3, produced a similar picture (Supplementary Fig. S11). The first difference, not surprisingly, is that the methods identify up to 20% less true positives. Another change (Supplementary Fig. S11b) is that under the alternative consideration of false positives (pairs from the same fold but different superfamilies, with DALI Z-score < 3 are ignored), the relative difference in sensitivity between the methods COMER and HHsearch becomes smaller. We observed from the analysis of the sensitivity and alignment quality that the inferred mixture model (2.2) particularly increased sensitivity to remote homology. This could account for the reduced difference between COMER and HHsearch when the definition for true positives is adjusted to closer homology.

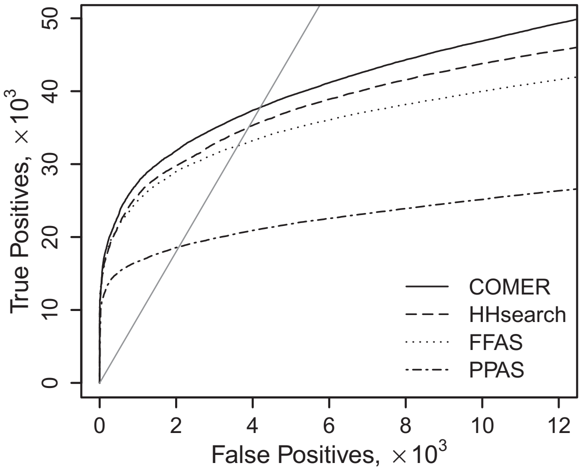

Fig. 1. True positives against false positives for the new method COMER and the other evaluated methods. Domain pairs belonging to the same SCOPe superfamily or sharing statistically significant structural similarity (DALI Z-score ≥ 2) are defined as true positives. Other pairs are false positives. The thin straight line represents a false discovery rate of 10%

We now shortly discuss PPAS lagging behind the other methods by a large margin in all the TP-vs.-FP plots. PPAS is a semi-global alignment method, which aligns two sequences over their entire length and is primarily designed for modeling whole proteins. Since (semi-)global alignments are well known to be less sensitive than local alignments, this could be one of the reasons of the lag of PPAS.

4.2 Alignment quality

A reference-free evaluation framework based on structural models (Margelevičius and Venclovas, 2010) was used to evaluate the quality of the profile alignments. According to this framework, an alignment between a pair of domains is considered to be of high quality if the most accurate of two 3D models generated for each domain (using the other domain as a template) is similar to the real structure.

We used MODELLER (Šali and Blundell, 1993) (version 9.4) to generate a 3D model based on a profile alignment and TM-score (Zhang and Skolnick, 2004) to evaluate the structural similarity between the model and the real structure. TM-score ranges from 0 to 1 with values ≥ 0.4 indicating a statistically significant similarity and hence implying a high-quality alignment. A profile alignment is assumed to be of low-quality if TM-score < 0.2, a characteristic value for a random pair. Alignments generating TM-score values between 0.2 and 0.4 are considered of neither high- nor low-quality and are ignored. Yet, to evaluate all top-ranked alignments, we provide a complementary analysis.

Importantly, the evaluation framework relies on two modes, *local* and *global*. The global mode evaluates the alignment with respect to the entire protein domain, whereas the local mode along the alignment extent. The global mode outweighs a performance gain in the local mode if profile comparison methods produce too short alignments, but it does not penalize over-extension of alignments. From the opposite position, the local mode counterbalances too long alignments tolerated by the global mode, but it allows under-extension of alignments. Hence, the analysis only under both modes provides a comprehensive evaluation of the quality of alignments.

Figure 2(A,B) shows the number of high-quality versus low-quality alignments in the local and global evaluation modes for top-ranked alignments sorted by their statistical significance. The new method COMER at a rate of low-quality alignments of 1% in the local mode produces 366% more high-quality alignments than PPAS, 181% than FFAS and 35% than HHsearch. In the global mode, at a rate of low-quality alignments of 10%, COMER produces 2% more high-quality alignments than PPAS and 11% than HHsearch, but 4% less than FFAS. Calculating unit-length contexts increased the rate of high-quality alignments for COMER by 10% and 2% in the local and global modes, respectively. Whereas, by computing 13−length contexts, HHsearch benefited from the 5% and 2% increase in the local and global modes, respectively.

The figure reveals all the methods competing in the global mode. In the local mode, as FFAS and PPAS generate long alignments, their performance falls behind. Trying to equalize the conditions for all the methods, we set up the methods COMER and HHsearch (its option -mact 0) to extend their alignments maximally (FFAS does not provide such a control). This set up does not affect the sensitivity. The results are shown in Figure 2(C,D). The picture is now opposite: The methods compete in the local mode (except PPAS) but in the global. In the global mode, at a rate of low-quality alignments of 10%, COMER produces 102% more high-quality alignments than PPAS, 90% than FFAS and 6% than HHsearch.

Figure 2 demonstrates that COMER compares favorably with the other methods when either local alignment or (semi-)global alignment over the entire protein length is emphasized. Nevertheless, we reconsidered the alignments produced by the methods to include into the analysis those inducing TM-score values between 0.2 and 0.4. Particularly, we were interested in the rate of high-quality alignments (TM-score ≥ 0.4) against all other alignments (TM-score < 0.4), although shifting the threshold TM-score to a value of 0.45 or 0.35 yielded similar results. The results obtained without and with setting up COMER and HHsearch to maximally extend their alignments (Figs S12 and S13 of the Supplementary Materials, respectively) are similar to those shown in Figure 2. However, in the local mode, COMER focused on local alignment (Supplementary Fig. S12a) outperforms the other methods by a smaller margin than in Figure 2A. It means that COMER still produces a rather large amount of alignments falling into the TM-score interval $[0.2; 0.4)$. This interval of TM-score values mainly corresponds to inter-fold relationships of proteins, some of which are remotely homologous or sharing local or global structural similarity. Figure 1 illustrates that some out of that amount of the COMER alignments represent those true positives.

4.2.1 Alignment quality of true positives

To examine the alignment quality of true positives and the impact on it of the model (2.2) describing profile contexts, we plot in Figure 3 the distribution of TM-scores in the local and global evaluation

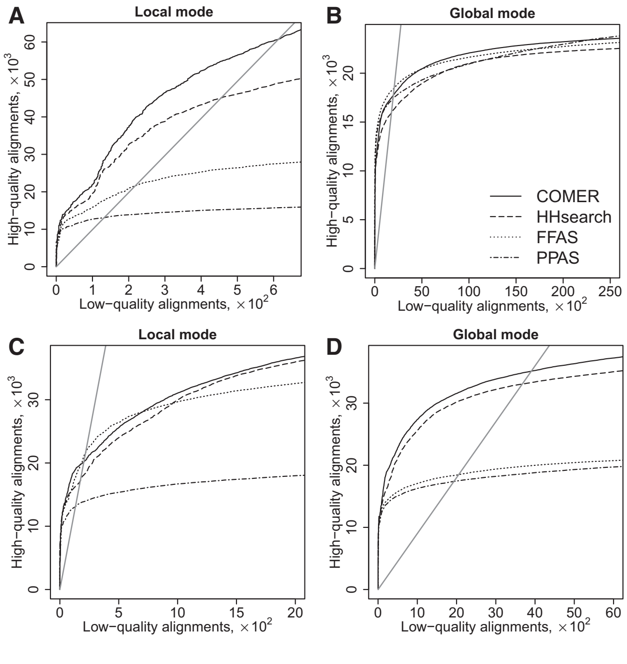

Fig. 2. Alignment quality of COMER and the other evaluated methods in the (A,C) local and (B,D) global evaluation modes. In (C) and (D), the methods COMER and HHsearch (-mact 0) maximally extend their alignments. A high-quality alignment is that producing a structural model of TM-score ≥ 0.4. TM-score < 0.2 corresponds to a low-quality alignment. Alignments generating $0.2 \leq$ TM-score < 0.4 are ignored. The thin straight line represents an error rate of 1% in the local mode (A,C) and 10% in the global mode (B,D). A nearly equal number of alignments are evaluated in both modes

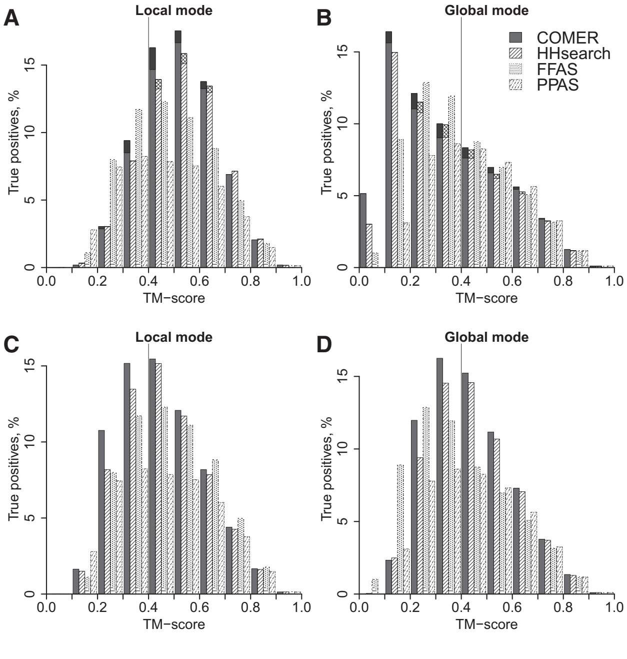

Fig. 3. Distribution of TM-scores for the true positives out of the top 90 000 alignments of each evaluated method in the (A,C) local and (B,D) global evaluation modes. In (A) and (B), dark-filled and cross-hatched boxes on top depict the percentage of true positives gained by calculating profile contexts by COMER and HHsearch, respectively. In (C) and (D), the methods COMER and HHsearch (-mact 0) were set up to extend their alignments maximally. The thin line at TM-score = 0.4 represents the limit of statistically significant structural similarity

modes for the true positives out of the top 90 000 alignments of each method. The definition of true positives matches that in Figure 1. The amount of true positives with a high-quality alignment (TM-score ≥ 0.4) the new method COMER produces in the local mode (Fig. 3A) constitutes 57% of the top alignments. This is 111% more than PPAS, 46% than FFAS and 8% than HHsearch. By calculating contexts, COMER increased the number of true positives attaining alignments of TM-score ≥ 0.4 and TM-score ≥ 0.2 by 3% and 4%, respectively (Fig. 3A and B). The corresponding increase for HHsearch is 2%.

Because of producing local alignment rather than (semi-)global, a large mass of the COMER alignments evaluated in the global mode (Fig. 3B) scores TM-score < 0.4. However, when COMER is set up to extend its alignments, in the global evaluation mode it yields 39% high-quality alignments, as illustrated in Figure 3D. This value is 50% greater than high-quality alignments produced by PPAS, 56% than that by FFAS and 5% than that by HHsearch (-mact 0). If we additionally take into the account the remotely homologous true positives that mainly come into the analysis at a lower statistical significance level and occupy the TM-score range $[0.2; 0.4)$, then COMER's improvement with respect to PPAS is 60% more true positives aligned with TM-score ≥ 0.2, 34% with respect to FFAS and 10% with respect to HHsearch. Narrowing down the definition of true positives by increasing the DALI Z-score threshold to 3 provides a similar COMER's improvement in alignment quality over the other methods in the local and global modes (Figs S14 and S15 of the Supplementary Materials).

4.3 Homology examples

COMER significantly detected a number of pairs of homologous protein domains that were not identified by the other three methods among their top 90 000 alignments. Here we provide two examples.

The first pair involves the chain (SCOP ID *d*1*rtwa*_) of putative transcriptional activator and that of the ribonucleotide reductase R2 (*d*1*mxra*_). Although the chains are representatives of different SCOP folds a.132 and a.25, respectively, clearly they share the same topology (Supplementary Fig. S16a). The homologous relationship between the two chains is confirmed by the ECOD evolutionary classification of protein domains (Cheng *et al.*, 2014). ECOD assigns them to the same homology group and also to the same topology group of Heme oxygenase/Ribonucleotide reductase.

The other example relates 2-keto-3-deoxy-6-phosphogluconate aldolase (*d*1*wa*3*a*_) to the uncharacterized protein YagE (*d*3*n*2*xa*_). SCOP (version 1.75) does not characterize the latter protein, but SCOPe categorizes both domains into the same superfamily of Aldolase, c.1.10. They exhibit the topology of TIM barrels (Supplementary Fig. S16b) and share the same homology and topology groups in the ECOD classification.

4.4 Evaluation on the basis of changed inputs

In this section, we evaluate the methods given a different set of MSAs that originated from searching the same domain sequences with HHblits (Remmert *et al.*, 2012) (version 2.0.16) using default settings for three iterations against the UniProt20 database of profile HMMs (06/2015). Having MSAs built on pairwise comparisons of profile HMMs allows one to examine how the methods perform when they operate with the more accurate input MSAs. We evaluated the sensitivity of each method and the quality of the alignments produced by each method in the same way as described above and summarize the results below.

4.4.1 Sensitivity

COMER at a false discovery rate of 10% (Supplementary Fig. S17) detects 84% more true positives than PPAS, 21% than FFAS and 14% than HHsearch.

4.4.2 Alignment quality

The more accurate MSAs caused a substantial increase in alignment accuracy for the methods (Supplementary Fig. S18). We thus sum up at the following rates of low-quality alignments. COMER at a rate of low-quality alignments of 0.2% in the local mode (Supplementary Fig. S18a) produces 300% more high-quality alignments than PPAS, 229% than FFAS and 60% than HHsearch. In the global mode having COMER and HHsearch set up to extend their alignments (Supplementary Fig. S18d), at a rate of low-quality alignments of 1%, COMER yields 101% more high-quality alignments than PPAS, 86% than FFAS and 20% than HHsearch.

The amount of true positives with a high-quality alignment (TM-score ≥ 0.4) COMER produces in the local mode (Supplementary Fig. S19a) constitutes 65% of the top 90 000 alignments. This is 99% more than PPAS, 57% than FFAS and 7% than HHsearch.

That amount in the global mode with the setup of extended alignments (Supplementary Fig. S19d) is 50% of the top 90 000 alignments. This is 59% more than PPAS, 63% than FFAS and 6% than HHsearch (-mact 0).

4.5 A test on CASP targets

Section 4 has heretofore described the evaluation of the methods based on the protein domains from the SCOPe database. Some domains, however, represent a part of the protein chain, and some of them may inadequately reflect practical situations of structure prediction. We thus tested COMER and the other methods on full-length sequences by a procedure similar to that applied for assessing structure prediction methods in CASP experiments.

We used all 85 CASP11 (2014) regular target sequences with available 3D structures. 3D models of the targets were generated using structural templates identified by each method and evaluated with TM-score. A database of templates corresponding to a release of the PDB database (Berman *et al.*, 2000) just before the start of CASP11 (05/01/2014), filtered to 20% sequence identity, equipped all the methods. Profiles for the target and template sequences were constructed according to the same operation as described in the beginning of Section 4, implying that the same set of MSAs supplied each method. The methods COMER and HHsearch were set up to extend alignments between profiles maximally.

The results obtained by searching the target profiles against the template profile database and evaluating target 3D models generated from produced profile alignments are shown in Table 1. The

Table 1. Average TM-score obtained by considering top *N* alignments per target for each method

	COMER	HHsearch	FFAS	PPAS
$N=1$	**0.488/0.533**	0.464/0.514	0.480/0.522	0.462/0.515
$N=2$	**0.466/0.499**	0.446/0.482	0.461/0.486	0.443/0.484
$N=3$	**0.451/0.475**	0.433/0.460	0.446/0.462	0.435/0.461
$N=4$	**0.441/0.452**	0.424/0.437	0.431/0.440	0.424/0.439
$N=5$	**0.428/0.428**	0.413/0.413	0.419/0.419	0.414/0.414

Numbers following forward slashes represent the average TM-scores achieved after reranking top 5 alignments for each target according to TM-score a posteriori. The highest values are highlighted in bold.

average TM-score of the COMER models is the largest, irrespective of the number of top alignments per target N (up to 5) considered. The differences between the average TM-score for COMER and those for the other methods are not large, but according to Wilcoxon signed-rank tests, the medians of the differences of the paired TM-scores for each number of top alignments considered deviate significantly from 0 at a confidence level of 95%. It means that the distributions of TM-scores obtained by using COMER differ from the distributions obtained by the other methods. In other words, the methods produce different alignments for the corresponding targets.

Plotting the distribution of TM-scores for COMER against that for another method reveals (Supplementary Fig. S20) some points corresponding to low TM-scores for COMER alignments but higher for other method's alignments. The reason is that each method ranks templates differently, and the same template can be ranked, for instance, first by one method and third by another. Indeed, the bivariate distributions of TM-scores become more condensed after sorting the top alignments of each method according to TM-score a posteriori (Supplementary Fig. S21). Several outliers (colored dark red in Supplementary Fig. S21) for which COMER scores low correspond to the same target T0852 for which many templates can be detected at a statistically significant level. COMER ranked more closely related templates below the top 5. Clearly, better ranking of templates could improve all the methods (see Table 1, Supplementary Figs S20 and S21), but this is a complex problem that can be addressed in many different ways and is beyond the scope of this study. Interestingly, after reranking the top alignments, the distribution of TM-scores for COMER still differs significantly from those for the other methods at the 95% confidence level, irrespective of $N = 1, \ldots, 5$, with the highest average quality of 3D models (Table 1).

In the end, we emphasize that this test does not unveil how the methods would have performed in CASP11. A system specifically designed for protein homology modeling would require a more elaborate computational pipeline covering regular update of data, inspection of the quality of profiles, selection of appropriate templates given profile alignments, modeling by combining multiple templates, and other routines. This test just demonstrates the relative performance of the methods under equal conditions on the targets for which a limited number of templates are available.

5 Discussion

Highly important in many fields of biomedical research, protein homology modeling relies on protein sequence alignments. Alignments of high quality stimulate discovery of evolutionary related matches and thus are key to sensitivity to homologous proteins and, consequently, homology modeling itself. In this study, we have addressed the problem of protein remote homology detection by improving the quality of alignments between protein sequence families represented by sequence profiles. Modeling profile contexts, fixed-length profile fragments representing the environment of sequence families, previously undertaken by Biegert and Söding (2009), has been engaged to achieve this goal.

We have proposed an HDP mixture model to describe the distribution of contexts, where a logistic-normal distribution is used to model substitution probabilities in profile contexts to capture dependencies between amino acids. We have inferred the HDP mixtures from a large set of profile context data by combining Gibbs sampling with Metropolis updates. The Metropolis updates exploited to overcome the inefficiency of Gibbs sampling were implemented by a split-merge algorithm (Jain and Neal, 2004), which we have adapted for the HDP mixture model. $K = 1260$ distinct context clusters have been obtained by applying a parallel MCMC sampling algorithm developed to speed up posterior inference. We have developed profile construction and comparison computer utilities and, with respect to the inferred mixtures, derived scores, which we integrated into an extensive profile-to-profile comparison.

Evaluating results by generating protein 3D models for each alignment between a pair of profiles and measuring the similarity between the model and the native structure, we have demonstrated that adding the scores increased the number of accurate 3D models, and consequently the number of accurate alignments, in both evaluation modes: with respect to the entire protein domain (global), key to protein structure prediction, and within the alignment boundaries (local), important to sensitivity. A benchmark of the new method COMER, available under the GNU GP License v3, with three other profile-to-profile comparison methods shows a significant improvement in both sensitivity and alignment quality, and an improvement from calculating unit-length contexts is twice that of the HHsearch method employing 13−length contexts.

Although this study confined to unit-length contexts, the HDP mixture model is applicable for longer contexts. The context consisting of S vectors of A substitution probabilities transformed by the inverse logistic transformation can be represented by an $A \times S$ matrix $\mathbf{X}$. The random matrix $\mathbf{X}$ will have a matrix variate normal distribution with mean matrix $\mathbf{M}$ and covariance matrices $\boldsymbol{\Sigma}$ $(A \times A)$ and $\boldsymbol{\Psi}$ $(S \times S)$, or equivalently, $\mathrm{vec}(\mathbf{X}^{\mathrm{T}}) \sim \mathcal{N}_{AS}(\mathrm{vec}(\mathbf{M}^{\mathrm{T}}), \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$. The mixture model (2.2) applied to $\boldsymbol{\eta} \equiv \mathrm{vec}(\mathbf{X}^{\mathrm{T}})$ will then describe the distribution of $S-$length contexts, where the conditional distributions and the MCMC sampling algorithms presented in the study remain valid. A model that allows for correlations not just between residues but also between context positions would impart much greater flexibility in modeling profile contexts and specificity in recognizing structural contexts. Integrated into comparison of profiles, the model would very likely increase the sensitivity and accuracy of alignments further. Accordingly, our further work will investigate how effective longer contexts under the HDP mixture model are.

Acknowledgements

The author is grateful to the Associate Editor and the reviewers for helpful suggestions.

Funding

This work was supported by the Research Council of Lithuania [grant number MIP-049/2013].

Conflict of Interest: none declared.

References

Aitchison,J. and Shen,S. (1980) Logistic-normal distributions: some properties and uses. *Biometrika*, **67**, 261–272.

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

Altschul,S. *et al.* (2009) PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.*, **37**, 815–824.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Biegert,A. and Söding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl. Acad. Sci. USA*, **106**, 3770–3775.

Blackwell,D. and MacQueen,J. (1973) Ferguson distributions via Pólya urn schemes. *Ann. Stat.*, **1**, 353–355.

Cheng,H. *et al.* (2014) ECOD: An evolutionary classification of protein domains. *PLOS Comput. Biol.*, **10**, e1003926.

Eddy,S. (2011) Accelerated profile HMM searches. *PLOS Comput. Biol.*, **7**, e1002195.

Ferguson,T. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Stat.*, **1**, 209–230.

Fox,N. *et al.* (2013) SCOPe: Structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.

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

Henikoff,S. and Henikoff,J. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Holm,L. *et al.* (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.

Jain,S. and Neal,R. (2004) A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comput. Graph. Stat.*, **13**, 158–182.

Jaroszewski,L. *et al.* (2011) FFAS server: novel features and applications. *Nucleic Acids Res.*, **39**, W38–W44.

Jones,D. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Khafizov,K. *et al.* (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proc. Natl. Acad. Sci. USA*, **111**, 3733–3738.

Kryshtafovych,A. *et al.* (2014) CASP10 results compared to those of previous CASP experiments. *Proteins*, **82**, 164–174.

Margelevičius,M. and Venclovas,Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. *BMC Bioinformatics*, **11**, 89.

Montelione,G. (2012) The Protein Structure Initiative: achievements and visions for the future. *F1000 Biol. Rep.*, **4**, 7.

Moult,J. *et al.* (2014) Critical assessment of methods of protein structure prediction (CASP) – round X. *Proteins*, **82**, 1–6.

Murzin,A. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nguyen,V. *et al.* (2013) Dirichlet mixtures, the Dirichlet process, and the structure of protein space. *J. Comput. Biol.*, **20**, 1–18.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Roy,A. *et al.* (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.

Sadreyev,R. *et al.* (2009) Discrete-continuous duality of protein structure space. *Curr. Opin. Struct. Biol.*, **19**, 321–328.

Schäffer,A. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Schwede,T. (2013) Protein modeling: What happened to the “protein structure gap”? *Structure*, **21**, 1531–1540.

Sjölander,K. *et al.* (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Söding,J. and Remmert,M. (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Curr. Opin. Struct. Biol.*, **21**, 404–411.

Teh,Y. *et al.* (2006) Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.*, **101**, 1566–1581.

The UniProt Consortium. (2014) Activities at the Universal protein resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.

Venclovas,Č and Margelevičius,M. (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins. *Proteins*, **77**, 81–88.

Venkatraman,E.S. (2000) A permutation test to compare receiver operating characteristic curves. *Biometrics*, **56**, 1134–1138.

Šali,A. and Blundell,T. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Yan,R. *et al.* (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci. Rep.*, **3**, 2619.

Yu,Y. and Altschul,S. (2005) The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, **21**, 902–911.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.