A Multimodal Approach for Protein Remote Homology Detection

Pietro Lovato, Alejandro Giorgetti, and Manuele Bicego

Abstract—Protein remote homology detection represents a crucial and challenging task in bioinformatics: even if effective methods appeared in recent years, in several cases a proper characterization of remote evolutionary correlation can not be derived. In such situations, it may be possible that information derived from other sources helps, provided that it is possible to properly integrate such (even partial) information into existing models. In this paper, we provide some evidence that this route is feasible: inspired by the multimodal retrieval literature, we show how it is possible to exploit a simple multimodal approach to improve a model learned from a set of sequences, by using knowledge derived from a partial set of corresponding 3D structures. We investigate (with the SCOP 1.53 benchmark) the suitability of the proposed multimodal scheme, showing that a beneficial effect can be obtained even when a very reduced amount of structures are available. A further detailed analysis on a member of the GPCR superfamily confirms that this multimodal approach can extract information that cannot be obtained from sequence-based techniques.

Index Terms—Multimodal approach, Ngrams, FragBag, topic models, GPCR

1 Introduction

PROTEIN homology detection is a central task in computational biology: it permits to identify functionally-related proteins, typically by looking at amino acid sequence similarity. For some homologue proteins this similarity may be low: in such cases, detecting the homology becomes a very challenging problem, typically referred to as *remote* homology detection. Many efficient approaches have been presented in the literature to face this problem [1], [2], [3], some of them based on discriminative methods such as Support Vector Machines [4], [5], [6], [7], [8], [9].

Even if reaching satisfactory accuracies on several benchmark datasets (e.g. the SCOP 1.53 dataset—[4]), there are still complex cases where even these state-of-the-art approaches may perform poorly. In such cases, it may be possible that information derived from other sources helps, provided that it is possible to properly integrate such (even partial) information into existing models. In the context of protein remote homology detection, there is a source of information which is typically disregarded by classical approaches: the available experimentally-solved, possibly *few*, 3D structures.[1] Now the question is: *Is it possible to improve sequence-based methods by integrating information derived from such 3D structures?* In this paper we provide some evidence that this is possible, by deriving a *multimodal* approach[2] for remote homology detection. We took inspiration from the multimodal image and text retrieval context [11], where images are equipped with loosely related narrative text descriptions, and retrieved by using textual queries. This scenario is particularly interesting with respect to our scopes, because it shares many similarities with our context: *i)* the link between the modalities is weak, partially hidden, and, in general, difficult to infer; *ii)* most importantly, the context is *asymmetric*: one of the two modalities is richer than the other, yet being more difficult or expensive to obtain—therefore fewer examples are typically available. The goal is to develop an approach which works directly on the weaker source of information (the text), being however built taking into account the (possibly smaller) richer source (the image).

In this paper we show that such multimodal point of view can be tailored to the protein remote homology detection case: in particular, the richer modality is represented by a (possibly small) subset of structures—retrieved from PDB—which are used to derive a "structure-aware" model for sequences. Our multimodal approach, based on the recent paper [12], starts by encoding sequences and structures with a count representation, namely a representation obtained by counting the number of occurrence of some basic elements inside an object: sequences are described using counts of Ngrams, as done in other effective protein remote homology detection approaches [6], [7], [13], whereas structures are described using counts of 3D fragments, as in [14]. Both representations are then modeled using topic models, a class of probabilistic approaches for count data: in particular we investigate here two models, the Latent Dirichlet Allocation (LDA) [15] and the Componential Counting Grids (CCG) model [12]. The former is a very famous topic model, recently employed also in this context [13], whereas the latter represents a recent and advanced admixture model which enriches topic models with topological constraints (its use in the protein remote homology detection context has never been investigated).

For both models, we created an *augmented model* accounting for structural information in two steps: *i)* a model (LDA or CCG) for the available structures is learned, creating a latent space which acts as a common, intermediate representation; *ii)* all the sequences are embedded into this space derived from structures. Such embedding is determined by exploiting the (partial) available correspondences between sequences and structures.

The suitability of the proposed multimodal framework for protein remote homology detection has been evaluated in two ways: on one hand, we performed various tests on the standard SCOP 1.53 benchmark [4], demonstrating that *i)* the proposed framework permits drastic improvements in those scenarios where sequence modality fails—even when only 10 percent of *training* sequences have their corresponding structure; *ii)* on the whole benchmark (54 families), it favorably compares with other recent approaches. On the other hand, we performed a thorough analysis on a member of the GPCR superfamily, suggesting that the proposed multimodal approach can extract information that cannot be derived by employing only sequence-based approaches.

2 Background

This section briefly summarizes the two probabilistic models employed in our approach, which belong to the wide family of "topic models" [16]. Topic models have been originally introduced in the text analysis community, in order to describe and model a set of documents. The basic idea underlying these methods is that each document may be characterized by the presence of one or more hidden topics (e.g. sports, finance, politics), inducing the presence of some particular words. From a probabilistic point of view, the document is then a mixture of topics, each one providing a probability distribution over words.

To employ these models, documents should be represented with an occurrence matrix (count matrix), where each entry $n^t(w_i)$ *counts* the number of times a given word w_i occurs in a given document (indexed by t). In our biological scenario documents

1. Some papers already show the potentialities of using structural information (see for example [10]); however, they are all based on 3D predictions made from sequences, therefore not using the true 3D structures found in PDB.

2. From a general point of view, a multimodal approach represents a technique aimed at solving a given task by integrating different sources of information.

- *P. Lovato and M. Bicego are with the Department of Computer Science, University of Verona, Verona, Italy. E-mail: {pietro.lovato, manuele.bicego}@univr.it.*
- *A. Giorgetti is with the Department of Biotechnology, University of Verona, Verona, Italy. E-mail: alejandro.giorgetti@univr.it.*

Manuscript received 6 Nov. 2014; revised 25 Mar. 2015; accepted 10 Apr. 2015. Date of publication 19 Apr. 2015; date of current version 5 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2424417

correspond to proteins, while basic building blocks (such as sequence Ngrams) are the observed words. Once learned, the topic models permit to represent all proteins in the topic space: even if in the protein case this space does not have a straightforward biological meaning,[3] it turned out to be really informative for comparing proteins, as largely shown in [18]. In the following, we will present the two topic models investigated, namely the Latent Dirichlet Allocation (LDA, [15]—perhaps the most famous topic model) and a recent extension called Componential Counting Grid (CCG, [12]).

2.1 Latent Dirichlet Allocation

Given a set of V different words, the LDA mediates the observation of a particular word w_i in a document t through a latent topic variable $z, z \in Z = \{z_1, \ldots, z_Z\}$, which is picked from a multinomial distribution $p(z \mid t) = \theta^t$. The multinomial θ^t represents the topic proportions, peculiar for every document t:

$$\begin{aligned} p(w_i^t) &= p(\theta^t \mid \alpha) \sum_k p(z_k|\theta^t) p(w_i^t|z_k) \\ &= p(\theta^t \mid \alpha) \sum_k \theta^t_{z_k} \cdot \beta_{w_i,z_k}. \end{aligned} \tag{1}$$

The topic z_k represents a probabilistic co-occurrence of words encoded by the distribution $p(w_i|z_k) = \beta_{w_i,z_k}$. Intuitively, θ^t measures the level of presence of each topic in the document t. On the other hand, β_{w_i,z_k} expresses how much a word w_i is related to the topic z_k. Finally, $p(\theta^t \mid \alpha)$ is a Dirichlet prior over the possible topics' assignments.

As better detailed in [15], the various distributions of the model are learned using a variational Expectation-Maximization (EM), a technique that maximizes the log-likelihood (or its tractable lower bound called Free Energy) by iterating between two steps: the E-step, which computes the posterior over the topics (i.e., θ^t), given the current estimate of the model; the M-step, where the parameters of the models (α and β) are re-estimated, given the current θ^t. Once the model has been trained, it is possible to use the learned parameters α and β to perform inference, estimating topic proportion $\theta^{t_{\text{new}}}$ of an unseen document t_{new}.

2.2 Componential Counting Grid

The Componential Counting Grid (CCG—[12]), introduced in the context of text mining, is a recent extension of LDA. The model stems from the fact that for many text corpora, documents *evolve* into one another in a smooth way, with some words dropping and new ones being introduced. For example, news stories smoothly change across the days, as certain evolving stories progressively fall out of novelty and new events create new stories. CCG introduces these topological constraints by arranging topics in a two-dimensional grid; topics, represented as *windows* inside the grid, may overlap in neighboring positions of the grid. More formally, the componential counting grid is a grid of discrete locations $\pi_{x,y}$, with fixed dimensions $\mathbf{E} = E_1 \times E_2$. Each location is endowed with a distribution over all V words, which acts exactly like the distribution β for LDA: given a location $z_k, k = (x, y)$ (i.e., a topic), π_k represents a multinomial distribution describing the probability of each word given that location (i.e., a topic). To model smooth transitions between topics, CCG assume that a word is not generated from a single distribution π_k related to a single position of the grid k (as in LDA), but also considering distributions in a neighborhood of k. In particular, a word in a document t is generated by i) choosing a location z_k from a multinomial distribution $p(z \mid t) = \theta^t$ (like topics proportion of LDA); ii) sampling from the average of all the π_k relative to a window of fixed dimensions $\mathbf{W} = W_1 \times W_2$ centered at z_k.

3. In some other cases—like the gene expression context [17]—a biological interpretation can be easily assigned.

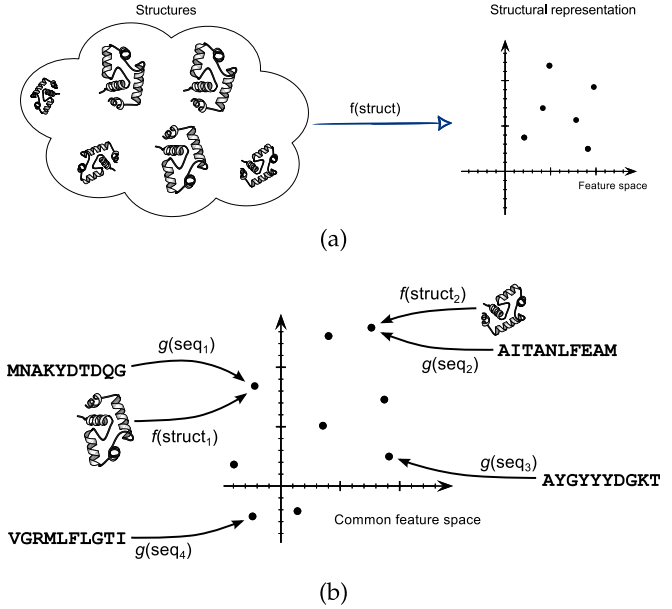

Fig. 1. The idea of the multimodal scheme.

As detailed in [12], model parameters and hidden distributions are learned using a variational EM algorithm. Similarly to LDA, the model is completely specified given the parameters α (Dirichlet prior over locations) and π. Again, given these quantities, inference on an unknown object permits to recover the value of $\theta^{t_{\text{new}}}$.

3 The Proposed Approach

In this section the multimodal approach used to integrate structural and sequential information is explained. From a very general perspective, the main idea is the following (see Fig. 1): suppose we have a set of sequences $\{seq_i\}$; for some of them we also know the corresponding structures $\{struct_i\}$. Then, we determine, from the set of structures $\{struct_i\}$, a function $f(struct)$, which is able to project all structures in a feature space (Fig. 1a). The goal is to determine a function $g(seq)$ so that $f(struct_i) \equiv g(seq_i)$ for all available structures (i.e., corresponding sequences and structures should share the same representation). The found function f can now be used to project whatever sequence in the common space, which is now built using structural information (Fig. 1b).

In order to realize this, we exploit an approach derived from the multimodal image-text retrieval literature [12], which is based on topic models described in the previous section. Even if different alternatives exist [11], [19], in such retrieval context the approach proposed in [12] appeared to be simpler and more effective.

3.1 Data Representation

Topic models assume that documents (proteins, in our case) are represented as *counting vectors*. Given a *dictionary* containing all possible words $w_i, i = 1, \ldots, V$, an entry in the count vector $n^t(w_i)$ represents the number of occurrences in the document t of the ith word of the dictionary. In our case, we need a counting representation for both sequences and structures. For the sequence modality, we use as words the so called Ngrams (i.e., short sequences of consecutive amino acids). Despite its simplicity, this representation has been already successfully exploited by other protein remote homology detection approaches [6], [7], [13]. In particular, in all our experiments we used bigrams, i.e., fragments composed by two consecutive amino acids. In the structural domain, we employed as words structural fragments, as proposed in [14]: each fragment is a list of 3D coordinates for consecutive Cα atoms in the backbone of the protein—in their original work, the authors provide different dictionaries of fragments. In our study, following other papers [14],

[18], we employed the 400_11 dictionary (composed by 400 structural fragments each of length 11).

At the end, we have two different dictionaries, one for each modality: a dictionary $D^{ST} = \{w_1^{ST}, \ldots, w_{V_{ST}}^{ST}\}$ for structures, and a dictionary $D^{SE} = \{w_1^{SE}, \ldots, w_{V_{SE}}^{SE}\}$ for sequences.

The input of our method is composed by:

- A set containing S pairs of corresponding sequence/structure counts, for a subset of training proteins

$$\{(ST_{Tr}^t, SE_{Tr}^t)\}, t = 1, \ldots, S,$$

where

$$ST_{Tr}^t = n^t(w_i^{ST}), i = 1 \ldots, V_{ST}$$
$$SE_{Tr}^t = n^t(w_i^{SE}), i = 1, \ldots, V_{SE}.$$

- A set of $T - S$ sequence counts, representing sequences in the training set without the corresponding 3D structure

$$\{SE_{Tr}^{S+1}, \ldots, SE_{Tr}^T\}.$$

- A set of N testing sequences

$$\{SE_{Te}^1, \ldots SE_{Te}^N\},$$

where $SE_{Te}^t = n^t(w_i^{SE})$.

3.2 Multimodal Learning

The key idea of the proposed multimodal approach is that the latent topic space learned by LDA (or CCG) establishes a common representation where both sequences and structures can be embedded. Since the two modalities are *asymmetric* (with the structural being the richer one), we impose this latent space to be powered by (possibly few) structures. The proposed approach articulates in three major steps:

Topic model learning on structures. First of all, we learn a topic model (LDA or CCG) using the available structure counts $\{ST_{Tr}^1 \ldots ST_{Tr}^S\}$: acknowledged the superiority of the structural modality, we emphasize the topic space to be "structure-driven".

For what concerns the learning, it is known that choosing a good initialization for parameters β (π for CCG) is crucial for a proper learning. Typically, this is done at random, with the risk of solution convergence to poor local minima. In order to overcome this issue, in our approach we perform a careful initialization: in particular, we cluster words into Z groups (where Z represents the number of topics) using the complete link algorithm, which performs an agglomerative clustering. Then, we initialize β (π) so that each topic has high probability of generating the words inside its cluster, and low probability of generating words outside the cluster.

At the end of this learning stage, each structure is characterized in such space by its corresponding vector $\theta_{ST}^t, t = 1, \ldots, S$.

Multimodal projection. In this step, we exploit correspondences between structures and sequences, projecting the sequences in the latent space learned with structures in the previous step. We impose that the topic proportions θ_{SE}^t for the S training sequences are equal to the θ_{ST}^t obtained from the corresponding structures. In this way we are establishing a 1:1 mapping between the structural topics and the sequential topics. In practice, this is achieved by learning the LDA/CCG model on sequence counts keeping θ_{SE}^t fixed and set to θ_{ST}^t. As a result, the parameters β_{SE} and α_{SE} (π_{SE} and α_{SE} for CCG) of the learned model are completely specified in the sequence domain. However, they have been learned taking into consideration the topic proportions derived from the model learned on structures.

Inference on the remaining training and testing sequences. For training proteins in the set $\{SE_{Tr}^{S+1}, \ldots, SE_{Tr}^T\}$, where 3D structures are unknown, an inference step with the learned enriched model can be performed to recover the topic proportions $\theta_{SE}^t, t = S+1, \ldots, T$. The same inference is performed on testing sequences to derive θ_{SE}^t for $SE_{Te}^t, t = 1, \ldots, N$. As explained in the background section, inference is performed by keeping fixed α, β (α and π for CCG), and estimating θ_{SE}^t for the new samples.

3.3 Classification Scheme

n order to perform classification, we employed a so-called generative embedding scheme [20], where the learned topic models are exploited to map the objects to be classified into a feature space, where a discriminative classifier can be used. Indeed, the topic posterior θ^t is a feature vector—already proven to be effective in several scientific fields [12], [21], [22], [23]—which can be used to train a discriminative classifier such as an SVM. SVMs are therefore trained using all θ_{SE}^t ($t = 1, \ldots, T$) in the training set, whereas classification is carried out on θ_{SE}^t ($t = 1, \ldots, N$).

4 Experimental Evaluation

In this section the proposed approach is evaluated with the standard and widely used SCOP 1.53 benchmark [4]. In particular, we first perform a thorough analysis on two cases where it is evident that the sole sequence modality fails, showing that drastic improvements can be obtained by the multimodal approach, even if using few structures; then we evaluate the proposed approach on the whole benchmark, in order to have a clear comparison with alternative approaches in the state of the art.

Both analyses are based on the SCOP 1.53 dataset [4], a famous benchmark widely employed to assess the detection capabilities of many PRHD systems. Such dataset, extracted from SCOP version 1.53[4] [24], contains 4,352 sequences from 54 different families. For each family, class labels are very unbalanced, with a vast majority of objects belonging to the negative class. Detection accuracies are typically measured using the receiver operating characteristic (ROC) score [25], which represents the area under the ROC curve (the larger this value the better the detection).

4.1 First Analysis: Families 3.42.1.1 and 3.42.1.5

In this first part we performed a thorough analysis on two cases where the sequence modality fails (i.e., cases where a proper characterization of the family cannot be determined). In particular, we concentrate on families 3.42.1.1 and 3.42.1.5, on which almost random accuracies are obtained by using models based on the sole sequences. We applied the proposed multimodal scheme on these two families, starting from the corresponding 3D structures downloaded from PDB. In particular, once encoded the sequences and the structures as explained in previous sections, the models (LDA or CCG) are learned from the training set, in order to get the θs usable to train the SVM. θs for the testing set are then extracted via model inference. When using LDA (and in general topic models), the number of topics should be set in advance, this representing a classic model selection problem (different solutions exist, such as hold-out likelihood, cross-validation, or, more in general, a priori knowledge). In this first analysis, taking inspiration from [6], [18], we set it to 100. For CCG, we exploited the concept of *capacity* [12], which measures how many non-overlapping windows can fit onto the grid. This can be assimilated to the number of topics in a topic model: therefore we set the CCG dimension as $E = [20, 20]$ and $W = [2, 2]$, so that the capacity equals to 100. After computing the θs, the classification has been carried out using the public libsvm implementation[5] [26], employing the RBF kernel. Parameter C of the SVM has been set as 10^{-3} for every experiment, whereas the RBF parameter σ has been found by exhaustive search, retaining for each family the one performing better on average (reasonable values lie around 2^{-2}).

4. http://noble.gs.washington.edu/proj/svm-pairwise/
5. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

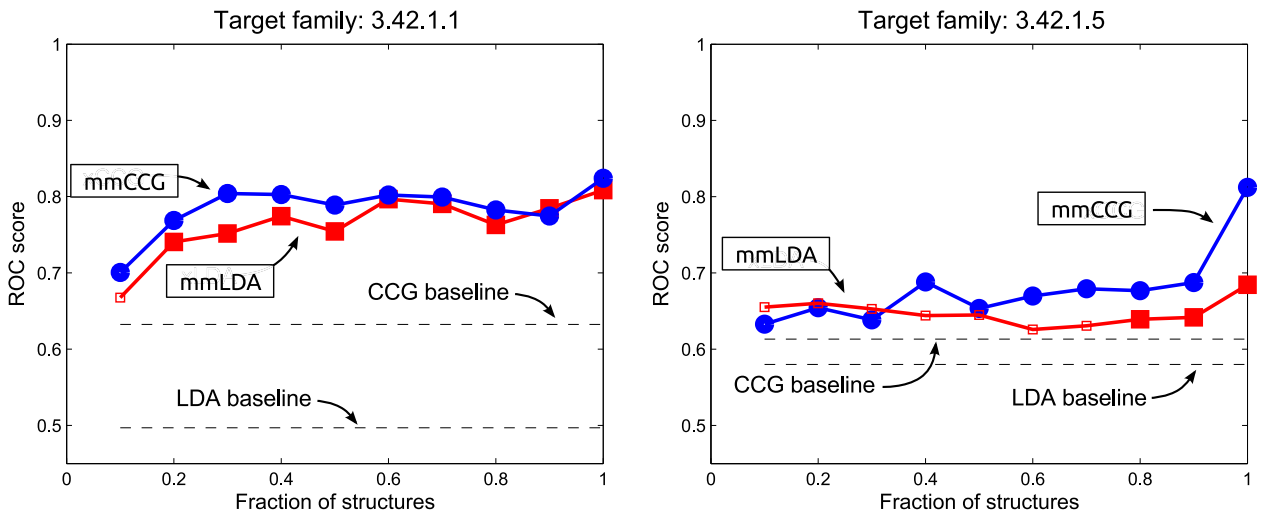

Fig. 2. Detection scores displayed as a function of the number of structures used in the multimodal approach. "mmLDA" ("mmCCG") stands for the proposed multimodal approach by using the LDA (CCG) model. Filled markers indicate statistically significant improvements over the baseline. Results are reported for (left) family 3.42.1.1 and (right) family 3.42.1.5.

In order to get a complete understanding of the proposed approach, we also assessed the performances when only a limited number of structures are available for learning. In particular, we used an increasing fraction of randomly chosen structures to build the structure model. Since there is a very limited number of positive examples (29 for the first family, 26 for the second), we decided to always consider all of them, sampling at random negative training examples. The structure model is then transferred to sequence model; inference on the enriched sequence model finally permitted to get descriptors for all training and testing sequences, to be used by the SVM classifier. Detection results, for fractions ranging from 0.1 to 1 (i.e., all training structures), are averaged over 50 runs, and reported in Fig. 2, for both the LDA and CCG models. We also determined whether the improvement gained with the proposed multimodal approach is statistically significant, using a standard t-test with alternative hypothesis "multimodal results are greater than the baseline". In Fig. 2, filled markers indicate statistical significance with p-value lower than $\alpha = 0.05$.

From these plots it seems evident that the use of structural information permits to derive a better sequence model: in both families, CCG achieves significant improvements when employing only 10 percent of all training structures. For the second family, even if multimodal LDA accuracies are higher than the baseline, statistical significance is obtained only when 80 percent or more of the structures are employed. When all training structures are considered, the improvement is rather high for both models.

When comparing the two probabilistic models, it appears evident that the Componential Counting Grid outperforms the LDA model, both when used on the sequence modality alone and when employed in a multi modal framework. Such a model, never used in the context of protein remote homology detection, permits to derive a better and more discriminant description of count data, confirming the results outlined in [12] for other application fields.

4.2 Second Analysis: All Families

In this second analysis, the proposed approach has been tested on all the families of the SCOP dataset, this being particularly important to compare the proposed scheme with the state of the art. In this case we slightly changed some details of our experimental pipeline; in particular, since we are dealing with 54 different classification problems (i.e., 54 families), we did not fix a single number of topics, but we let it vary in a reasonable range, keeping the best value. Moreover, in order to be fully comparable with many works in the state of the art [6], [7], [8], [9], [27], [28], the classification is performed using SVM via the public GIST implementation,[6] setting the kernel type to radial basis, and keeping the remaining parameters to their default values.

Results are presented in Table 1, in comparison with the literature; in particular, the state of the art is split into methods which employ Ngrams (*Ngram-based Methods*) and methods which do not (*Other Methods*). From the table it can be observed that the framework is rather accurate: when compared with other Ngram-based methods, our best result outperforms almost all other approaches, the only exception being the SVM-Top-Ngram-combine [7] approach, for which an almost equivalent detection rate was reported. In such approach, however, different Ngram representations are combined: in order to completely demonstrate the potentialities of our proposed approach, we followed a similar idea, by combining different representations extracted from the multimodal CCG model. The result is presented in Table 1 as "Multimodal Combined CCG", and clearly confirms that margins of improvements are still present. From the table, it is also interesting to consider that the proposed multimodal technique compares reasonably well also with other more complex approaches. Finally, interestingly CCG outperforms LDA only when used in a multi modal framework.

5 Multimodal Analysis of Bitter Taste Receptor TAS2R38

The main goal of this section is to qualitatively validate the proposed multimodal scheme in a real scenario. In particular we focus on a specific protein (the bitter taste receptor TAS2R38 [29], [30]) belonging to the G-protein coupled receptors (GPCRs) superfamily. This large group (with over 900 members only in humans) of cell signaling membrane proteins is of major importance for drug development, as GPCRs are one of the primary targets currently under investigation [31].

From our perspective, this context is very interesting for three reasons: *i)* sequence identities between members of different GPCR families are extremely low, making the detection of remote homologues very challenging; *ii)* only 24 unique human GPCRs[7] have their experimentally-determined structure as of January 2015 (i.e., very little structural information); *iii)* most importantly, it has already been shown that the closest homologue of the TAS2R38

6. Downloadable from http://www.chibi.ubc.ca/gist/ [4]

7. The list of such proteins is obtained from http://blanco.biomol.uci.edu/mpstruc/

TABLE 1
Average ROC Scores for the 54 Families in the SCOP 1.53 Superfamily Benchmark for Different Methods

Method	ROC	Reference
Monomodal LDA	0.921	This paper
Monomodal CCG	0.903	This paper
Multimodal LDA	0.925	This paper
Multimodal CCG	0.932	This paper
Multimodal Combined CCG	0.941	This paper
Ngram-based methods		
SVM-Ngram	0.826	[6]
SVM-Ngram-LSA	0.878	[6]
SVM-Top-Ngram (n=1)	0.907	[7]
SVM-Top-Ngram (n=2)	0.923	[7]
SVM-Top-Ngram-combine	0.933	[7]
SVM-Ngram-p1	0.887	[9]
SVM-Ngram-KTA	0.892	[9]
Other methods		
SVM-pairwise	0.896	[5]
SVM-LA	0.925	[5]
SVM-Pattern-LSA	0.879	[6]
SVM-Motif-LSA	0.860	[6]
PSI-BLAST	0.676	[6]
Profile *(5,7.5)*	0.980	[27]
SVM-Bprofile	0.921	[28]
SVM-PDT-profile (β=8,n=2)	0.950	[8]
HHSearch	0.915	[8]
SVM-LA-p1	0.958	[9]

receptor (as given by standard programs for sequence search, without manual intervention) does not represent a good template usable to unravel structural/functional elements (in particular, regarding the active site and the specific residues involved in the ligand binding) [32]. We show here that our multimodal approach can be used to suggest an alternative template. We sponsor this template by providing some elements supporting the capabilities of the obtained multi modal model of capturing structural/functional elements. To do that, a multimodal LDA (with three topics[8]) has been trained, using all sequences and the known 24 structures (downloaded from PDB): as a result, all GPCR sequences are embedded in the topic probabilities θ space. The query TAS2R38 sequence is embedded in the same space via inference on the model: the nearest neighbor with known structure represents the suggested template. In this case we have the N/OFQ Opioid Receptor (PDB id: 4EA3). On the contrary, if we perform the same analysis with the single modality LDA, we obtain as nearest neighbor the CCR5 chemokine receptor (PDB id: 4MBS); as described above, modeling TAS2R38 using this template alone does not allow a correct characterization of the binding cavity of the receptor [32].

To validate the new template, we try to mine the obtained multimodal model, in order to see if the contained information exhibits structure-driven importance. To do that, we analyze, for every topic, the five most probable Ngrams (as given by β distribution), trying to understand if they are related to positions in the two proteins which are important from a structural point of view. Actually we have found that some of these Ngrams (shown in the top part of Fig. 3, together with the topic probabilities θ of the query and the corresponding nearest neighbor) represent words which are located with primary importance in the binding cavity of both proteins—these critical residues already shown to be involved in ligand recognition on our query TAS2R38 [33]. If we repeat the same analysis using a LDA model built using only sequences (central part of Fig. 3), no evident structural or functional information can be derived, this preliminary suggesting that the N/OFQ Opioid

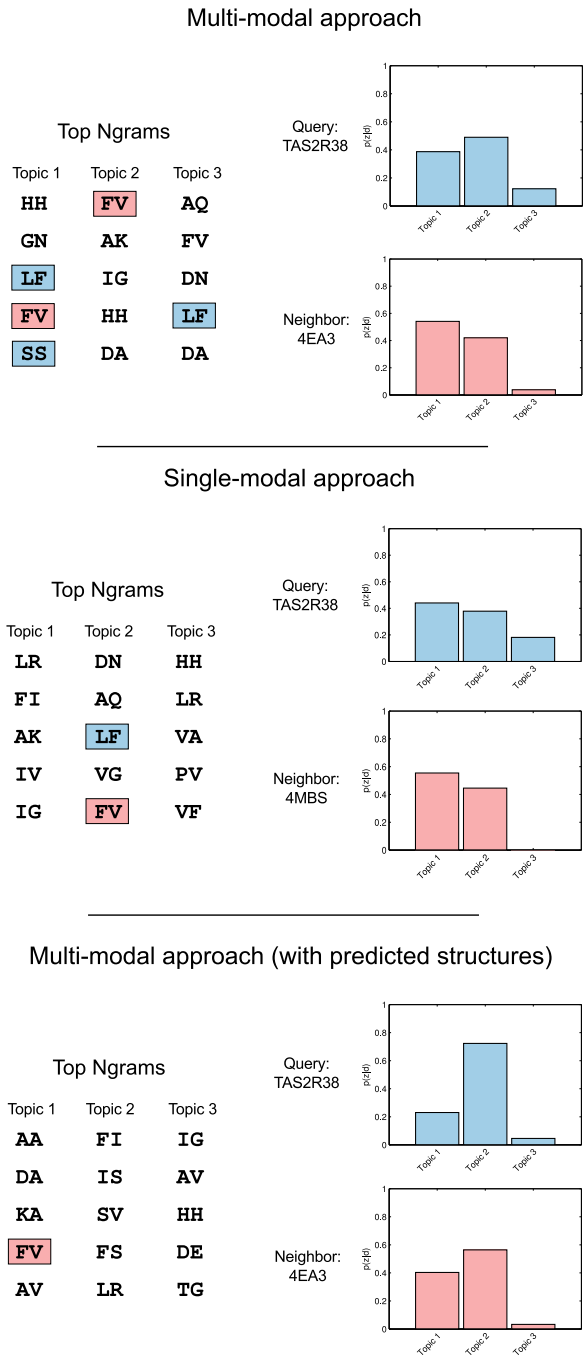

Fig. 3. On the top part of the figure, the first five Ngrams (sorted in descending order w.r.t their β probabilities) for each topic are listed. Ngrams highlighted are known to occur in the binding site locations of either of the two proteins. Slightly to the right, θ distributions (with three topics) are displayed for the query TAS2R38 and its closest neighbor. In the central part of the figure, we visualize the same information employing the LDA in a single-modal way. Finally, in the bottom part of the figure, the same information has been extracted with the multimodal approach employing both real and predicted structures. Interestingly, adding such predicted structures deteriorates the qualitative results obtained by the multimodal scheme.

Receptor, being obtained with a more "structure aware" model, can represent a valid alternative to the CCR5 chemokine receptor.

A final experiment has been carried out in to investigate if it may be possible, in cases like this when very few structures are available, to enlarge the structural information of the training set by also using predicted 3D structure models.[9] To test this we

8. In this case we had to drastically reduce the number of topics since only 24 structures are available—the topic space is built by using the structural information.

9. For example those obtained using http://zhanglab.ccmb.med.umich.edu/GPCR-HGmod/

applied our proposed multimodal approach by enlarging the training set with the predicted structures of different proteins belonging to the TAS2R group (24 GPCR models, downloaded from http://zhanglab.ccmb.med.umich.edu/GPCR-HGmod/). Results are displayed in the bottom part of Fig. 3: even if we obtain the same suggested template (the N/OFQ Opioid Receptor—PDB id: 4EA3), the quality of the multimodal space seems worst than that of the true multimodal approach. It seems that adding predicted models does not help the proposed approach, but, on the contrary, adds some noise. This was somehow expected, and confirms the intuition we got from the other quantitative experiments: the fully exploitation of the proposed framework is based on the use of a small piece of information, which should be however extremely informative (as is for real structures compared to simulated structures).

In conclusion, the availability of a method that, augmenting the descriptive power of a sequence-based model, is able to predict relevant structural positions, i.e., involved in ligand binding, is a fundamental step for setting up the modeling protocol when no 3D experimental information is available. In the studied case, the information obtained using our approach could be essential for guiding the selection of better and biologically relevant target-template alignments.

6 Conclusion

This paper investigated a multimodal approach for protein remote homology detection. In particular we provided some evidence that it is possible to improve sequence based models by exploiting the available (even partial) 3D structures. The approach, based on topic models, allowed the derivation of a common and intermediate feature space—the topic space—which embeds sequences being at the same time "structure aware". We experimentally demonstrated that, in cases where the sequence modality alone fails, introducing only 10 percent of the *training* structures resulted in significant improvements on detection scores. Moreover, we applied the proposed approach to model a GPCR protein, finding evidences of structural correlations between sequence Ngrams: such correlations can not be recovered employing a sequence-only technique.

As a final consideration, we would like to point out that this multimodal scheme seems to be particularly suitable for those situations where the sequence modality fails (as shown in our quantitative and qualitative experiments). When the sequence modality is already performing adequately, the improvements are not so significant: probably in such cases the simple scheme we investigated in this preliminary work (which simply postulates the equivalence of the structure and the sequence spaces) is not flexible enough to significantly improve the results. We are currently studying more robust multimodal approaches, which can for example learn how to move from the structure space to the sequence space.

Acknowledgments

P. Lovato is the corresponding author.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psiblast: A new generation of protein database search programs," *Nucleic Acid Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[2] K. Karplus, C. Barrett, and R. Hughey, "Hidden Markov models for detecting remote protein homologies," *Bioinformatics*, vol. 14, pp. 846–856, 1998.

[3] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.

[4] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *J. Comput. Biol.*, vol. 10, no. 6, pp. 857–868, 2003.

[5] H. Saigo, J.-P. Vert, N. Ueda, and T. Akutsu, "Protein homology detection using string alignment kernels," *Bioinformatics*, vol. 20, no. 11, pp. 1682–1689, 2004.

[6] Q. Dong, X. Wang, and L. Lin, "Application of latent semantic analysis to protein remote homology detection," *Bioinformatics*, vol. 22, no. 3, pp. 285–290, 2006.

[7] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining top-n-grams and latent semantic analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 510, 2008.

[8] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, p. e46633, Sep. 2012.

[9] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, and K.-C. Chou, "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[10] Y. Hou, W. Hsu, M.-L. Lee, and C. Bystroff, "Efficient remote homology detection using local structure," *Bioinformatics*, vol. 19, no. 17, pp. 2294–2301, 2003.

[11] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. International Conf. Comput. Vis.*, 2011, pp. 2407–2414 .

[12] A. Perina, N. Jojic, M. Bicego, and A. Truski, "Documents as multiple overlapping windows into grids of counts," in *Proc. Adv. Neural Inform. Process. Syst. 26*, 2013, pp.10–18.

[13] J. Yeh and C. Chen, "Protein remote homology detection based on latent topic vector model," in *Proc. Int. Conf. Netw. Inf. Technol.*, Jun. 2010, pp. 456–460.

[14] I. Budowski-Tal, Y. Nov, and R. Kolodny, "FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately," in *Proc. Nat. Acad. Sci.*, vol. 107, no. 8, pp. 3481–3486, 2010.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[16] D. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[17] M. Bicego, P. Lovato, A. Ferrarini, and M. Delledonne, "Biclustering of expression microarray data with topic models," in *Proc. Int. Conf. Pattern Recognit.*, 2010, pp. 2728–2731.

[18] S. Shivashankar, S. Srivathsan, B. Ravindran, and A. V. Tendulkar, "Multi-view methods for protein structure comparison using latent dirichlet allocation," *Bioinformatics*, vol. 27, no. 13, pp. 161–168, 2011.

[19] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. and Develop. Inf. Retrieval*, 2003, pp. 127–134.

[20] J. A. Lasserre, C. M. Bishop, and T. P. Minka, "Principled hybrids of generative and discriminative models," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 87–94.

[21] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in Proc. *Eur. Conf. Comput. Vis.*, 2006, pp. 517–530.

[22] M. Cristani, A. Perina, U. Castellani, and V. Murino, "Geo-located image analysis using latent representations," in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2008, pp. 1–8.

[23] M. Bicego, P. Lovato, A. Perina, M. Fasoli, M. Delledonne, M. Pezzotti, A. Polverari, and V. Murino, "Investigating topic models' capabilities in expression microarray data classification," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 6, pp. 1831–1836, Nov.-Dec. 2012.

[24] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "Scop database in 2004: Refinements integrate structure and sequence family data," *Nucleic Acids Res.*, vol. 32, pp. 226–229, 2004.

[25] M. Gribskov and N. L. Robinson, "Use of receiver operating characteristic (roc) analysis to evaluate sequence matching," *Comput. Chemistry*, vol. 20, no. 1, pp. 25–33, 1996.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011.

[27] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, 2005.

[28] Q. Dong, L. Lin, and X. Wang, "Protein remote homology detection based on binary profiles," in *Proc. 1st Int. Conf. Bioinformatics Res. Develop.*, 2007, vol. 4414, pp. 212–223.

[29] U.-k. Kim, E. Jorgenson, H. Coon, M. Leppert, N. Risch, and D. Drayna, "Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide," *Science*, vol. 299, no. 5610, pp. 1221–1225, 2003.

[30] B. Bufe, P. A. Breslin, C. Kuhn, D. R. Reed, C. D. Tharp, J. P. Slack, U.-K. Kim, D. Drayna, and W. Meyerhof, "The molecular basis of individual differences in phenylthiocarbamide and propylthiouracil bitterness perception," *Current Biol.*, vol. 15, no. 4, pp. 322–327, 2005.

[31] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?" *Nature Rev. Drug Discovery*, vol. 5, no. 12, pp. 993–996, 2006.

[32] X. Biarnés, A. Marchiori, A. Giorgetti, C. Lanzara, P. Gasparini, P. Carloni, S. Born, A. Brockhoff, M. Behrens, and W. Meyerhof, "Insights into the binding of phenyltiocarbamide (ptc) agonist to its target human tas2r38 bitter receptor," *PLoS ONE*, vol. 5, no. 8,p. e12394, 2010.

[33] A. Marchiori, L. Capece, A. Giorgetti, P. Gasparini, M. Behrens, P. Carloni, and W. Meyerhof, "Coarse-grained/molecular mechanics of the tas2r38 bitter taste receptor: Experimentally-validated detailed structural prediction of agonist binding," *PLoS ONE*, vol. 8, no. 5, p. e64675, 2013.