EVOLUTIONARY SEQUENCE MODELING FOR DISCOVERY OF PEPTIDE HORMONES

Kemal Sönmez *Lawrence Toll* *Nina Zaveri*

SRI International, Menlo Park, CA

ABSTRACT

We describe a computational framework that models spatial structure along the genomic sequence simultaneously with the temporal evolutionary path structure and show how such models can be used to discover new functional molecules through cross-genomic sequence comparisons. The framework incorporates a priori high-level knowledge of structural and evolutionary constraints in terms of a hierarchical grammar of evolutionary probabilistic models. In particular, we demonstrate a novel computational method for identifying novel prohormones and the processed peptide sites by producing sequence alignments across many species at the functional-element level. We present experimental results with an initial implementation of the algorithm used to identify potential prohormones by comparing the human and mouse proteins, resulting in high accuracy identification in a known set of proteins and a putative novel hormone from an unknown set. Finally, in order to validate the computational methodology, we present the basic molecular biological characterization of the novel putative peptide hormone, including identification in the brain and regional localizations. The success of this approach will have a great impact on our understanding of GPCRs and associated pathways, and help us identify new targets for drug development.

Index Terms— evolutionary HMM, peptide hormone, hierarchical grammar

1. INTRODUCTION

Evolution provides a vast array of clues for the discovery of functional proteins encoded in the genome. The genomic sequences, depending on the functionality of the structures they encode, get preserved or diverge in nonuniform ways across species genome [1, 2]. As the number of complete genomes builds up, a more complete picture of how evolution has operated on the principal biological structures emerges. Hence, biological sequence analysis techniques that are informed by phylogenetics in modeling specific functional structures become more compelling approaches to understanding the biological systems involved [3]. This paper describes a computational framework that models *spatial* structure along the genomic sequence simultaneously with the *temporal* evolutionary path structure. We demonstrate that this framework can be used to explore the dependencies among the genomes of several species and discover new functional molecules, in particular prohormones and the associated peptide hormones. The computational framework, which we name *HIGHER: HIerarchical Grammar of Hidden markov models of Evolutionary Regions*, uses probabilistic models that incorporate *a priori* high-level knowledge of structural and evolutionary constraints in terms of a probabilistic grammar of evolutionary HMM modules, which, in turn, model the low-level sequence homologies. The resulting cross-genomic models can be used in a generative or discriminative manner for modeling and alignment of sequences or detection of new molecules, respectively.

The paper is organized as follows. We first give some brief background on peptide hormones and computational methods for modeling peptides in the introduction. Then we introduce the hierarchical grammars of functional elements in Section 2. We present experimental results in Section 3, and conclude with a brief discussion in Section 4.

1.1 Peptide Hormones

Peptide hormones represent a large class of first messengers in a signal transduction pathway that operates through activation of G protein coupled receptors (GPCRs). GPCRs represent the largest gene family, making up perhaps 3% of the mammalian genome [2]. Because of their extracellular sites of action and importance as first messengers for cellular signaling, GPCRs have become a primary target for drug development. In fact, about half of all pharmaceuticals act as agonists or antagonists of GPCRs. For many newly identified GPCRs, an endogenous ligand is not known (orphan GPCRs), and discovering their ligands plays a key role in the discovery of drug candidates that act on the associated signal transduction pathways. Although the receptors possess an easily recognized primary and tertiary structure, their endogenous ligands, and the prohormones from which they are derived, are very difficult to identify. The set of known peptide ligands consists of short protein sequences and displays

few signs of homology or easily identifiable subdomains. The principal method today for identifying novel peptide hormones entails purification and sequencing of the active ligand from some mammalian tissue based upon activation of a known or orphan GPCR, which is a slow and costly process.

1.2 Computational Methods of Peptide Modeling

The problem of computational peptide hormone discovery based on the genomic structure alone proves to be difficult. An attempt to build models by specifying rules via deterministic grammars within the inductive logic grammar framework is described in [4], where, by enforcing the existence of signal sequences and splicing sites through a deterministic context-free grammar, a sieve for possible prohormone sequences is proposed. Even without the insight provided by evolutionary forces, the resulting method is able to eliminate structurally unlikely candidates, but due to the ubiquitous existence of double basic residues throughout protein sequences, its selectivity turns out to be poor.

Phylogenetic HMMs, or phylo-HMMs, are probabilistic models that combine HMMs and phylogenetic models in order to explain the spatial (genomic) and temporal (evolutionary) characteristics of a sequence. The first introduction of phylo-HMMs was motivated by the need to improve phylogenetic models that allow for variation in the substitution rate across sites [5, 6]. The problem of secondary structure prediction was addressed next [7, 8]. There has been a recent marked increase in the interest in these models as cross-genomic data become available in large quantities and approaches that are informed by evolutionary pressures become enormously useful [1, 9-12]. Particularly, they have been applied to cross-genome gene prediction [13, 14]. Another similar structure is the evolutionary HMM [15, 16] (note that we use this name more generally in this paper, not referring to this specific model only) that accounts for the phylogenetic information using generalizations of pairwise-HMMs, in a way similar to our approach. Evolutionary HMMs do not model the genomic structure directly, though, and the spatial part of the model is used to track the shifts in phylogenetic parameters. Match profile HMMs (MPHMMs) [17] combine the capabilities of two types of HMMs in that they use a profile HMM structure in modeling the sequence structure and a pairwise HMM (or a multiple-genome generalization) in modeling the evolutionary characteristics across species.

2. HIERARCHICAL GRAMMARS OF EVOLUTIONARY HMMS

Hierarchical grammars of evolutionary HMMs, such as phylo-HMMs or MPHMMs, are probabilistic models that take into account the way substitutions take place in the evolutionary path at specific sites along the genome and the specific patterns of change from one site to the next. Figure 1 shows a hierarchical grammar of evolutionary HMM modules for a prohormone. At the functional-level hierarchy, the model is specified in terms of its functional elements, such as signal sequences, splicing sites, and so on. The underlying evolutionary HMM modules carry out the local multiple alignments with respect to the phylogenetic relationship warranted by the context. This kind of hierarchical alignment is significantly more informative than a conventional multiple sequence alignment in that it provides a segmentation. For the hormone problem, the most important feature of a cross-genome alignment turns out to be the difference between the substitution rates of the functional and the nonfunctional subsequences around (predominantly double basic residue) splicing sites.

Let us define the computational structure of a hierarchical grammar of functional-evolutionary model modules (MPHMMs or phylo-HMMs) by the four-tuple, $\theta = (\mathbf{\Pi}, \mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ where $\mathbf{\Pi} = \{\pi_1, ..., \pi_M\}$ is a set of functional component states (for functions such as a signal sequence, a splicing site, or a peptide) with the set of associated functional element models, $\mathbf{G} = \{G_1, ..., G_M\}$, with the model G_j accounting for the part of the sequence alignment at the component state π_j. $\boldsymbol{\alpha} = \{\alpha_{jk}\}, (1 \le j, k \le M)$, and $\boldsymbol{\beta} = \{\beta_1, ...\beta_M\}$ are the matrix of component state transition probabilities and the vector of initial probabilities, respectively. In this formulation, for the sake of descriptive efficiency, we are describing the basic two-level hierarchy of models, which can, in our implementation, entail more levels. In the lower level of the hierarchy, each component

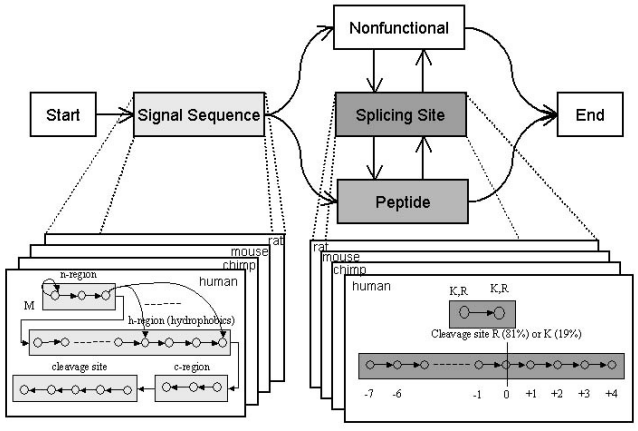

Figure 1. Prohormone hierarchical grammar and the evolutionary HMM modules.

model is a vector output HMM (such as a MPHMM) with an alphabet consisting of the four-tuple, $G_j = (\mathbf{S}^j, \boldsymbol{\lambda}^j, \mathbf{A}^j, \mathbf{b}^j)$, where S_j is a set of states associated with the functional component module. For example, a simple double basic residue cleavage site HMM would have two states that emanate multiple alignments of Arg and Lys residues. This structure also supports hierarchical grammars of phylo-HMMs [3]. In that case, $G_j = (\mathbf{Q}^j, \boldsymbol{\omega}^j, \boldsymbol{\sigma}^j, \boldsymbol{\delta}^j)$, where $\mathbf{Q}^j$ is the substitution matrix defined with respect to the alphabet of amino acids, $\boldsymbol{\omega}^j$ is a vector of equilibrium frequencies, $\boldsymbol{\sigma}^j$ is the binary phylogenetic tree with the set of branch lengths $\boldsymbol{\delta}^j$. Felsenstein's "pruning" algorithm [18] is used for the phylogenetic model optimization.

In this two-level hierarchical approach, there are two types of alignments, (i) functional alignments at the high level, $C = (C_1, ..., C_L)$, and (ii) state module alignments at the lower level, $X^k = (X_1^k, ..., X_{L_k}^k)$, $k = 1, ..., L$. To illustrate this point, Figure 2 shows a hierarchical alignment for Prepronociceptin from five species (human, chimp, mouse, rat, cow), where the boxes show the functional element sequence and actual sequence alignments are shown. Given the above setting, HIGHER computes the joint probability of a functional level path and alignment, which is given by

$$P(\phi, \mathbf{C} \mid \boldsymbol{\theta}) = \beta_{\phi_1} P(C_1 \mid G_{\phi_1}) \prod_{i=2}^{L} \alpha_{\phi_{i-1}\phi_i} P(C_i \mid G_{\phi_i}),$$

where, in turn, each of the functional module state alignments is given by

$$P(\boldsymbol{\tau}, \mathbf{X}^j \mid G_j) = b_{\tau_1}^j P(X_1^j \mid \lambda_{\tau_1}^j) \prod_{i=2}^{L} a_{\tau_{i-1}\tau_i}^j P(X_i^j \mid \lambda_{\tau_i}^j).$$

The likelihood of the model $P(\mathbf{C} \mid \boldsymbol{\theta}) = \sum_{\phi} P(\phi, \mathbf{C} \mid \boldsymbol{\theta})$ is found by summing over all possible paths, and the maximum likelihood path is the path that maximizes that sum. The computation of these quantities and the state posterior probabilities is facilitated by the Markovian structure that allows standard dynamic programming based solutions through the use of Viterbi and forward-backward algorithms.

The most compelling feature of the proposed computational framework is the enabling of the scientist to incorporate a priori functional-level knowledge directly into the model topology in a straightforward manner. Through its high-level grammar, it allows modeling and testing of hypotheses that are specified in terms of functional components such as signal sequences, splicing sites, and peptide hormones. Rather than merely exploratory analyses of the cross-genomic relationships, the hierarchical grammar of MPHMMs enables the biologist to specify a genomic structure along with its phylogenetic attributes.

3. EXPERIMENTAL RESULTS

3.1. Bioinformatics Results

As a proof of principle, we present results on SwissProt, a database containing a large number of known hormones. Because the functions of all the proteins in SwissProt are known, this search does not produce novel peptide hormones, but it produces a detection metric for the performance of the search paradigm. Note that the structural profile HMMs for the signal sequence and the splicing sites have not been trained with these proteins, and in HIGHER we do not train sequence structure models for hormones, so our SwissProt set constitutes an independent test set. For one specific threshold, we were able to identify 45 out of 54 prohormones known to be in SwissProt with 44 false alarms In terms of detection performance, this corresponds to a point on the ROC curve with sensitivity of 83%, and specificity of more than 99.9% (44 false hits on a SwissProt set with 122,564 proteins).

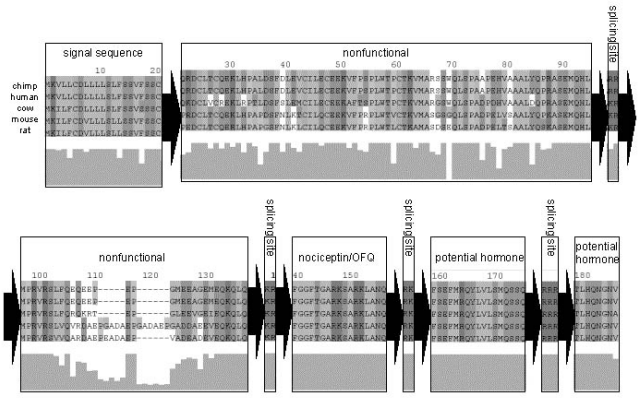

Figure 2. Hierarchical functional-element multiple alignment of Pronociceptin across human, chimpanzee, mouse. rat, and cow.

3.2. Biochemistry Results

We present preliminary biochemistry results on a protein (found across human and mouse proteins) that has so far been our top candidate. We believe that this will represent at least one novel neuropeptide, and we have named it Neuropeptide Q (NPQ), because three out of the four potential neuropeptides from this protein have at least one glutamine. This protein is the perfect example of our hypothetic neuropeptide model. Between double basic residues, the homology is high. Outside these residues the conservation is quite low. The protein sequence of the human and rat were predicted from gene

finding programs. These proteins have no apparent homology to any other proteins, and no known function.

Our preliminary investigation of this protein, using RT-PCR, shows transcripts in human, mouse, and rat brain. We have cloned and verified the sequence of the human, mouse, and rat cDNA. Northern analysis using a human tissue blot (Clontech) showed the presence of message in brain, pancreas, but most prominently in the kidney Therefore, NPQ may be one of many peptides (such as vasopressin) found in both brain and kidney. We have cloned and verified the sequence of the human, mouse, and rat cDNA. In collaboration with Dr. Stan Watson (University of Michigan) we have also conducted studies to determine regional localization in brain by in situ hybridization. This experiment shows a very discreet localization in what appears to be locus ceuruleus (LC). This peptide is therefore likely to be found in noradrenergic cells of the LC. These cells have projections in many parts of the brain and are likely to have effects on mental health and psychiatric diseases. We will soon know if this protein is processed in the way that we anticipate. The GPCR to which this peptide (or any of the peptides from this protein) binds is not known at this time. No other information is known about this protein and its potential peptides. In fact, other than computationally, there would be no reason to believe that this protein is processed in any special way. However, in light of our computational model of a neuropeptide and species comparison, it seems highly likely that there will be peptides generated from this protein that have biological activity.

4. CONCLUSIONS

We have presented a computational framework that is capable of accounting for protein structure and cross-species evolutionary divergence simultaneously. By aligning low-level evolutionary HMM modules within a high-level functional-element grammar, it is possible to build precise models of the effects of evolutionary pressures on genomic structures. In particular, we have applied this technique to modeling of prohormones across species with the goal of identifying novel prohormones and associated peptide hormones based on their evolutionary divergence profiles and genomic structures. The technique has resulted in high accuracy detection on a known dataset and led to putative hormones on a set of hypothetical proteins. Biochemical validation of the findings have started and produced promising results.

5. ACKNOWLEDGMENT

This work was supported by the NIH (NIDA) CEBRA program.

6. REFERENCES

1. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.
2. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
3. Siepel, A. and D. Haussler, *Combining phylogenetic and hidden Markov models in biosequence analysis.* J Comput Biol, 2004. **11**(2-3): p. 413-28.
4. Bryant, C.H., Muggleton, S.H., Srinivasan, A.,Whittaker, A., Topp, S., Rawlings, C., *Are grammatical representations useful for learning from biological sequence data? - a case study.* Linkoping Electronic Articles in Computer and Information Science, 2001. **6(2001)**(13).
5. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach.* J Mol Evol, 1981. **17**(6): p. 368-76.
6. Yang, Z., *A space-time process model for the evolution of DNA sequences.* Genetics, 1995. **139**(2): p. 993-1005.
7. Goldman, N., J.L. Thorne, and D.T. Jones, *Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses.* J Mol Biol, 1996. **263**(2): p. 196-208.
8. Thorne, J.L., N. Goldman, and D.T. Jones, *Combining protein evolution and secondary structure.* Mol Biol Evol, 1996. **13**(5): p. 666-73.
9. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
10. Boffelli, D., et al., *Phylogenetic shadowing of primate sequences o find functional regions of the human genome.* Science, 2003. **299**(5611): p. 1391-4.
11. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.
12. Thomas, J.W., et al., *Comparative analyses of multi-species sequences from targeted genomic regions.* Nature, 2003. **424**(6950): p. 788-93.
13. Pedersen, J.S. and J. Hein, *Gene finding with a hidden Markov model of genome structure and evolution.* Bioinformatics, 2003. **19**(2): p. 219-27.
14. McAuliffe, J.D., L. Pachter, and M.I. Jordan, *Multiple-sequence functional annotation and the generalized hidden Markov phylogeny.* Bioinformatics, 2004. **20**(12): p. 1850-60.
15. Holmes, I., *Using guide trees to construct multiple-sequence evolutionary HMMs.* Bioinformatics, 2003. **19 Suppl 1**: p. i147-57.
16. Holmes, I. and W.J. Bruno, *Evolutionary HMMs: a Bayesian approach to multiple alignment.* Bioinformatics, 2001. **17**(9): p. 803-20.
17. Sonmez, K. and L. Toll. *A Novel Hidden Markov Model for Cross-Genome Discovery of Peptide Hormones*. in *GENSIPS 2005*. 2005. Newport, RI, USA.
18. Felsenstein, J., *Maximum-likelihood estimation of evolutionary trees from continuous characters.* Am J Hum Genet, 1973. **25**(5): p. 471-92.