Tópicos em Bioinformática – 2º/2010

Rodrigo Vargas Honorato – 7399151

Hidden Markov Models (HMM)

W226–W229 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki471

KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites

Hsien-Da Huang*, Tzong-Yi Lee, Shih-Wei Tzeng[1] and Jorng-Tzong Horng[1,2]

Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan, [1]Department of Life Science and [2]Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan

Received February 13, 2005; Revised and Accepted April 15, 2005

- Descreve um servidor online para a identificação de sítios kinase-especificos.
- Forma mais abundante de regulação celular. Justificado pela importância no controle celular
- Necessidade de desenvolver um esquema computacional que seja capaz de facilmente e eficientemente identificar os sítios de fosforilação e também qual tipo de kinase está envolvida
- Database com dados experimentais, os sítios fosforilados foram extraídos como positivos e os não fosforilados negativos.
- *Maximal depence decomposition (MDD)* para agrupar os sets em subgrupos; MDD é um processo recursivo que divide o set em subgrupos baseado na dependência local das sequencias.
- H*idden markov model* (pHMM). Os sítios de ligação são os estados onde há probabilidade de encontrar um dado resíduo.
- Para cada set kinase-especifico é selecionado o melhor modelo que será usado para para identificar o sitio se fosforilação dada uma sequencia do input pelo HMMsearch.
- Bit score dado pelo HMMER é que define o match.

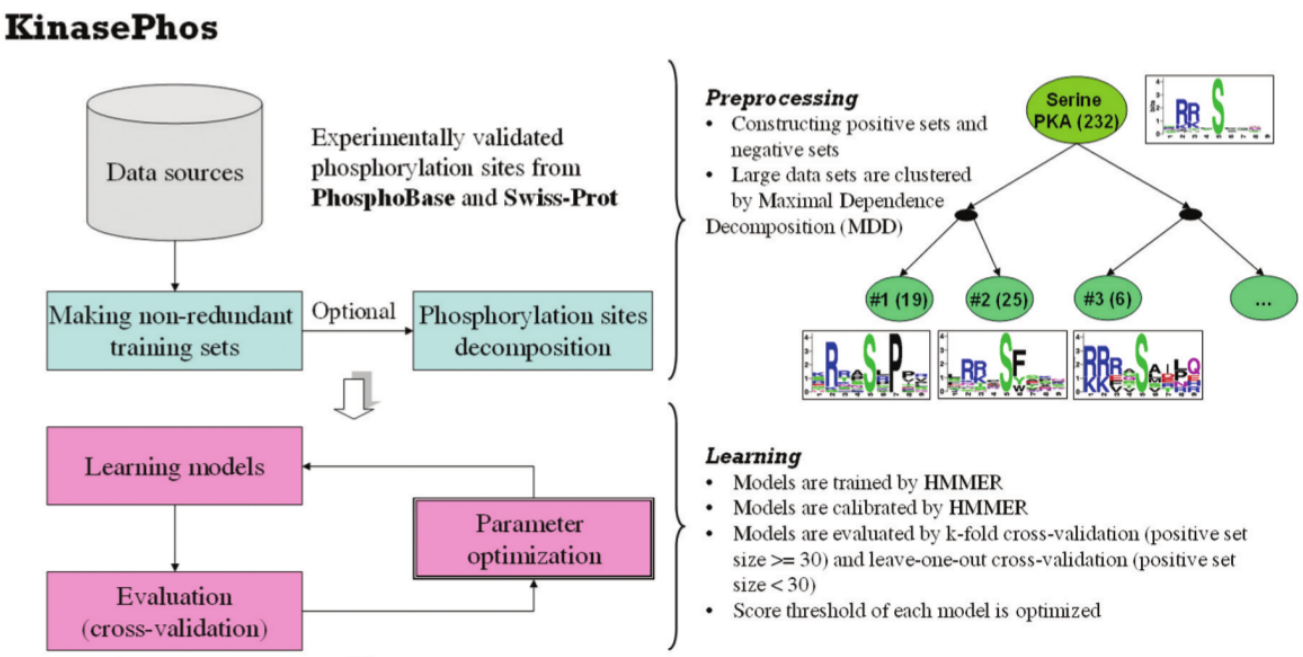

Figura 1 – Fluxograma mostrando o funcionamento do software.

BMC Structural Biology

BioMed Central

Research article

Open Access

Analysis of an optimal hidden Markov model for secondary structure prediction

Juliette Martin*[1,2], Jean-François Gibrat[2] and François Rodolphe[2]

Address: [1]INSERM U726, Equipe de Bioinformatique Génomique et Moléculaire Université Denis Diderot Paris 7, 2 place jussieu, 75251 Paris Cedex 05, France and [2]INRA, Unité Mathématiques Informatique et Génome, Domaine de Vilvert, 78352 Jouy en Josas Cedex, France

Email: Juliette Martin* - juliette.martin@jouy.inra.fr; Jean-François Gibrat - jean-francois.gibrat@jouy.inra.fr; François Rodolphe - francois.rodolphe@jouy.inra.fr

* Corresponding author

- Problema constante na biologia estrutural, a obtenção da estrutura secundária de uma proteína.
- Praticamente todos métodos de aprendizagem já foram usados pra tentar resolver esse problema e que o mais usado é o modelo de redes neurais.
- HMM já foram usados pra fazer essa predição.
- Novo HMM treinado a partir de estruturas secundárias sem conhecimento prévio, não levando em consideração conhecimento biológico.
- Cada estrutura secundária é modelada por um único estado, aqui existem diversos estados de emissão em cada estrutura secundária.
- Dificuldade está em achar o número ótimo de estados para cada estrutura secundária.
- Resolvido da seguinte forma:
  1. Considera-se modelos com um numero igual de estados para cada estrutura.
  2. É utilizando o escore Q_3 (número de resíduos na estrutura *i* preditos na estrutura *j*, dividido pelo número total de resíduos), o critério de informação Bayesiana, BIC = logL – 0.5 x K x log(N) o primeiro termo da função aumenta com o número de parâmetros e é penalizado pelo segundo e por ultimo pela distancia estatística entre dois modelos.
  3. Leva-se em consideração os modelos nos quais os critérios para somente um estado aumentam enquanto os outros permanecem fixos.

- Assim é definido o alcance do modelo a ser explorado para cada classe estrutural: entre

12 e 16 para hélices, 6 a 10 para folhas beta e 5 a 13 para ligações.

- 225 modelos que se encaixam nesses critérios são gerados e avaliados.
- O melhor modelo tem 36 estados, 15 para hélices, 9 para folhas beta e 12 para ligações mostrados na Figura 1.

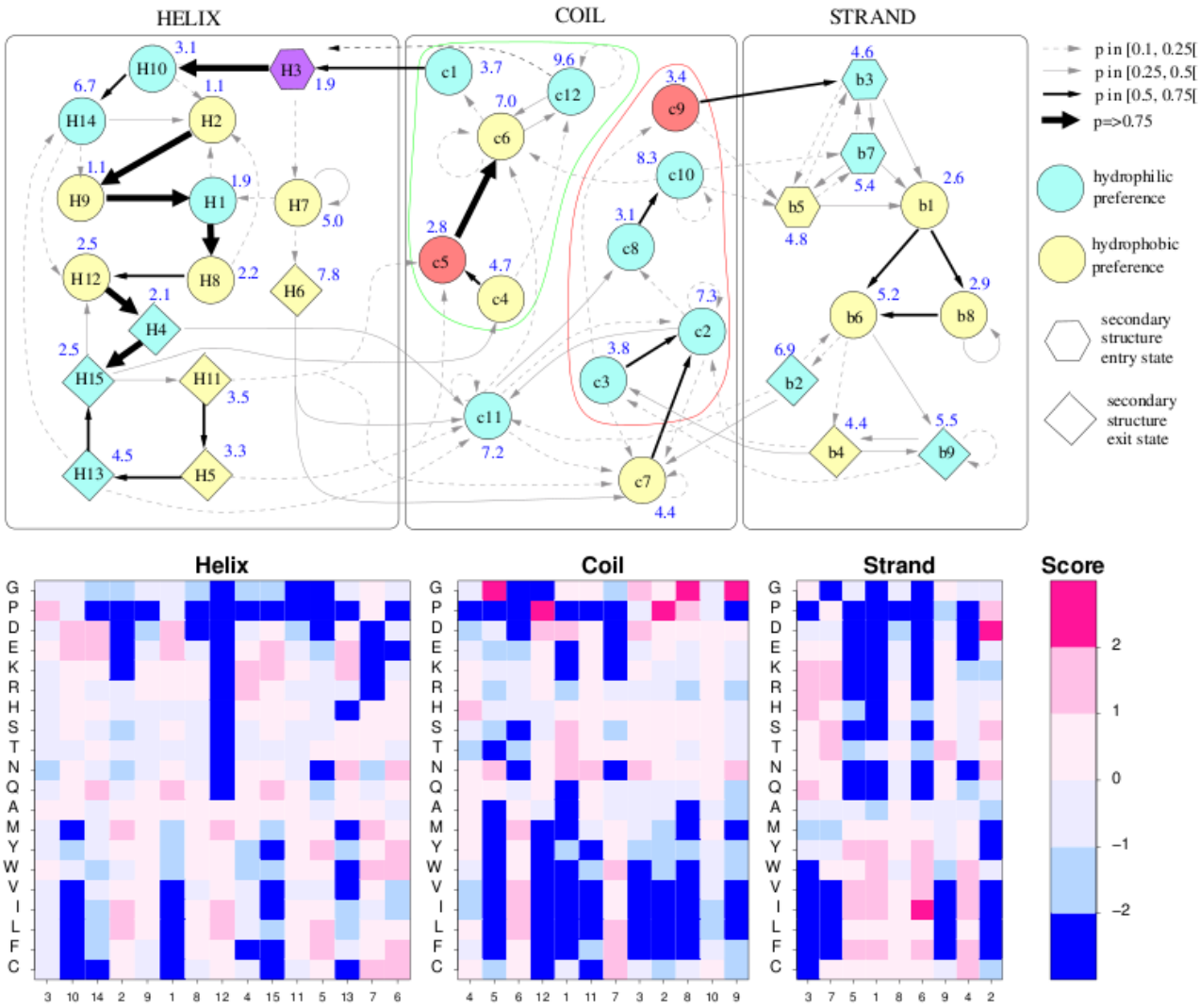

Figura 1 – Os 36 estados finais encontrados. Parte superior: somente transições com probabilidades associadas superiores a 0.1 são mostradas. A espessura das setas varia de acordo com o valor da probabilidade. Os estados são coloridos de acordo com a preferência por um determinada característica do aminoácido (hidrofóbico x hidrofílico), roxo indica que não há nenhuma forte preferência e vermelho o favorecimento de glicina. Parte inferior: propensão de cada aminoácido em cada um dos estados (P(a|s)) onde o escore igual a 1 indica que o aminoácido é duas vezes mais frequente naquele estado do que em todo o dataset.

- Valores em azul é o valor *Neq* derivado da função de entropia de Shannon $Neq(s) = exp[\sum - p(s; r) \ ln(p(s;r)) \ ]$, onde p(s;r) é a probabilidade da transição do estado *s* para o

estado *r*.

- Estruturação nas transições entre estados no quadro referente as hélices, isso é explicado graças ao padrão de formação de alfa-helices, resíduos hidrofílicos em contato com o solvente e hidrofóbicos.
- Presença de somente um estado de entrada para as hélices, o estado H3, que não favorece nem pune nenhum resíduo.
- O HMM é testado para prever a estrutura secundária de 505.
- Através de validação estatística chega-se a conclusão que o modelo é mais eficiente (escore Q_3 HMM: 67.9% e PSIPRED: 66.8%).
- Para múltiplas sequências chega-se a uma conclusão semelhante, onde os autores afirmam que apesar da simplicidade do modelo, ele parece promissor e talvez mais eficiente do que outros métodos que levam em consideração caracteres evolucionários.