Estymatory rozkładów prawdopodobieństwa występowania sekwencji aminokwasowej w uliniowieniu i ich zastosowanie do predykcji struktury trójwymiarowej białek

autoreferat rozprawy doktorskiej

Szymon Nowakowski

Maj, 2006

1 Wstęp

Niniejsza rozprawa doktorska zajmuje się tematyką leżącą na pograniczu trzech dziedzin wiedzy: biologii molekularnej, matematyki oraz informatyki. Sformalizowaliśmy w niej pojęcia związane z analizą sekwencyjną białek i DNA. Podajemy szereg matematycznych metod modelowania uliniowień, poruszamy również kwestię analizy jakości tych modeli, a także metod predykcji sekwencji dobrze do uliniowienia dopasowanych.

Motywacją dla przedstawionych w rozprawie badań jest zastosowanie prezentowanych wyników do predykcji struktury trójwymiarowej białek. Dr Krzysztof Fidelis zaproponował metodę predykcji struktury trójwymiarowej białek z użyciem lokalnych deskryptorów [17]. Omawiane w rozprawie techniki analizy uliniowień sekwencji są istotną częścią tej metody. Zagadnienie predykcji struktury białek jest jednym z najważniejszych we współczesnej bioinformatyce, gdyż możliwość znajdowania struktury białek *in silico*, czyli przy użyciu komputera, mogłaby wspomóc lub nawet wyprzeć niezwykle kosztowne (materialnie i czasowo) metody laboratoryjne, takie jak krystalografia rentgenowska oraz jądrowy rezonans magnetyczny. Poznanie struktury białka jest kluczowym momentem dla wielu badań biologii molekularnej i ma zastosowanie między innymi przy projektowaniu leków, produkcji detergentów i pestycydów, dla analizy funkcji nowych białek.

Weryfikacji istniejących technik predykcji struktury trójwymiarowej białek *in silico* służą przeprowadzane co dwa lata eskperymenty CASP (ang. *Critical Assessment of Protein Structure Prediction*) [1, 2, 3].

2 Uliniowienia sekwencyjne i ich opis

Uliniowienie można rozumieć jako prostokątną macierz P o K wierszach i M kolumnach. W klatkach macierzy stoją litery pewnego alfabetu $\mathcal{A}$. W rozprawie zajmujemy się głównie białkami i wtedy mamy na myśli alfabet dwudziestu liter, będących skrótami nazw aminokwasów, czyli

$$\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{K}, \mathtt{L}, \mathtt{M}, \mathtt{N}, \mathtt{P}, \mathtt{Q}, \mathtt{R}, \mathtt{S}, \mathtt{T}, \mathtt{V}, \mathtt{W}, \mathtt{Y}\}.$$

W pewnych rzadkich przypadkach będziemy mówić o alfabecie DNA, czyli

$$\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{T}, \mathtt{G}\}.$$

Wiersze macierzy P utożsamiamy z sekwencjami (odpowiednio — białkowymi lub DNA).

W rozprawie szczegółowo omawiamy stan obecnej wiedzy w dziedzinie estymowania na podstawie uliniowienia rozkładu prawdopodobieństwa określonego na przestrzeni sekwencji o ustalonej długości M. Rozkład estymuje się na podstawie uliniowienia traktowanego jako próba losowa sekwencji o długości M.

Obecnie podstawowym sposobem opisu uliniowień jest rozkład prawdopodobieństwa zwany PSSM (ang. *Position Specific Score Matrix*), w którym zakłada się niezależność kolumn uliniowienia. Jednakże w ostatnich latach pojawił się szereg prac [4, 5, 8, 16, 13, 15], w których analizowane są zależności, mogące pojawić się między kolumnami uliniowień.

W klasie rozkładów PSSM wiadomo, że najprostszy estymator (największej wiarogodności, NW) częstości występowania liter $\mathcal{A}$ w kolumnach uliniowienia nie jest wystarczający. Dla krótkich uliniowień może się bowiem zdarzyć, że jedna z liter w ogóle nie wystąpi w badanej kolumnie, czyli wyestymowana częstość wyniesie 0. Jest to niedopuszczalne w zastosowaniach biologicznych, w których staramy się unikać kategorycznych sądów, że coś nie jest możliwe (szczególnie dla mało licznych prób losowych). Dlatego w użyciu pojawiły się estymatory bayesowskie, umożliwiające wyspecyfikowanie wiedzy *a priori* o spodziewanej częstości występowania liter alfabetu $\mathcal{A}$. Do metod tych należą metody *zero-offset* [14], metoda psuedozliczeń (jeden rozkład Dirichleta jako wiedza *a priori*) [21], a także metoda oparta na mieszaninie wielu rozkładów Dirichleta [7, 20]. O tej ostatniej pokazano, że jest bliska teoretycznemu optimum w klasie metod PSSM [14].

Innym sposobem uniknięcia problemów pojawiających się przy estymacji NW jest użycie macierzy substytucyjnych. Działają one jednak optymalnie jedynie dla liczby wierszy $K = 1$ [14].

Istnieje szereg metod wprowadzenia zależności między kolumnami do opisu uliniowienia. Można zastosować ukryte modele Markowa (HMM, ang. *hidden Markov model*) [8], jednakże ten rodzaj opisu modeluje zależności jedynie między przyległymi kolumnami. Do modeli umożliwiających analizę odległych zależności należą między innymi mieszaniny PSSM, drzewa bayesowskie, a także mieszaniny drzew. Są to klasy sieci bayesowskich omawiane w kontekście bioinformatycznym w [5]. Inne stosowane metody polegają na analizie częstości krotek liter [13, 15].

3 Wyniki teoretyczne

3.1 Zaproponowane nowe metody opisu uliniowień

W rozprawie proponujemy nową metodę estymacji PSSM, która jest konkurencyjna do metody mieszaniny rozkładów Dirichleta, ale jest znaczenie szybsza obliczeniowo. Jest ona oparta na zastosowaniu wiedzy *a priori* w postaci rozkładów Beta.

Dowodzimy zgodność szerokiej klasy estymatorów modeli PSSM, między innymi udowodniona jest zgodność estymatora opartego o metodę rozkładów Beta.

Proponujemy też nowatorską metodę wprowadzenia zależności do modelu uliniowienia, która dopuszcza zastosowanie bardziej złożonej wiedzy *a priori*, niż do tej pory stosowane techniki. Przy metodach uwzględniających odległe zależności między kolumnami do tej pory stosowano technikę pseudozliczeń (jeden rozkład Dirichleta) jako wiedzę *a priori* [5], my zaś podajemy metodę estymacji mieszaniny rozkładów PSSM (czyli modelu z zależnościami) z mieszaniną wielu rozkładów Dirichleta jako wiedzą *a priori*. Do estymacji parametrów proponowanego modelu stosujemy estymator MAP (ang. *Maximum a Posteriori*). Używamy w tym celu algorytm EM (ang. *Expectation Maximization*) [9, 11, 12], służący do estymacji MAP przy obecności ukrytych zmiennych (między innymi w przypadku mieszanin). Użycie wielu rozkładów Dirichleta jest ważne, gdyż każdy rozkład Dirichleta w mieszaninie może modelować inną fizykochemiczną cechę stowarzyszoną z literami alfabetu [7, 20]. W przypadku aminokwasów można mówić o takich ich cechach, jak między innymi hydrofobowość, polarność, ładunek (pozytywny bądź ujemny), rozmiar, skłonność do występowania w pętlach.

3.2 Badanie statystycznej istotności predykcji i jakości modeli

W rozprawie poruszamy też temat badania statystycznej istotności predykcji (polegających na stwierdzeniu „badana sekwencja lub jej fragment dobrze pasuje do badanego uliniowienia"). Predykcje takie były do tej pory badane zazwyczaj przy użyciu p-wartości jako miary ich statystycznej istotności [5]. Wskazujemy

w rozprawie przyczyny tego, że p-wartość nie jest wystarczającą miarą i w szczególnych przypadkach może źle odpowiadać rzeczywistości. Proponujemy nową metodę, którą nazywamy metodą Powtórnej Weryfikacji [19], która na podstawie niezależnej próby pochodzącej z populacji sekwencji negatywnych (na podstawie kryteriów biologicznych do uliniowienia nie pasujących) oraz pozytywnych (dobrze pasujących do badanego uliniowienia) pozwala ocenić statystyczną istotność predykcji.

Ustalmy badane uliniowienie. Jeśli na niezależnych próbach sekwencji, o których wiadomo, że są pozytywne (negatywne) wyestymujemy wartości P^+, P^- oraz rozkłady $P(\cdot|+), P(\cdot|-)$, gdzie

- P^+, P^- to prawdopodobieństwa *a priori* tego, że badana sekwencja jest, odpowiednio, pozytywna lub negatywna, zanim zobaczymy samą sekwencję,
- $P(\cdot|+), P(\cdot|-)$ to rozkłady badanej oceny predykcji (np. oceny *log-odds* [10]) otrzymanej przez sekwencje z populacji, odpowiednio, pozytywnej lub negatywnej.

Wtedy ocenę Powtórnej Weryfikacji dla wartości oceny predykcji s można sformułować z użyciem twierdzenia Bayesa jako

$$\mathrm{PW}(s) = P(+|s) = \frac{P(s|+)P^+}{P(s|+)P^+ + P(s|-)P^-}.$$

Kolejną kwestią, którą rozważamy w rozprawie, jest ocena jakości modelu dla uliniowienia. Za dobrej jakości uznamy model, w którym dla sekwencji pozytywnych (dobrze pasujących do badanego uliniowienia) otrzymamy wysokie oceny w sensie Powtórnej Weryfikacji. Jeśli w badanym modelu dla sekwencji, o których wiemy, że są pozytywne, dostaniemy niskie oceny, to taki model uważamy za słabej jakości. Intuicja ta została sformalizowana w postaci Oczekiwanej Powtórnej Weryfikacji [19] — miary zbudowanej na podstawie Powtórnej Weryfikacji (jako wartość oczekiwana Powtórnej Weryfikacji w rozkładzie pozytywnym):

$$\mathrm{OPW} = E^+(\mathrm{PW}) = \int_{-\infty}^{\infty} \frac{P^2(s|+)P^+\,ds}{P(s|+)P^+ + P(s|-)P^-}.$$

Wiąże ona z modelem probabilistycznym uliniowienia liczbę z przedziału $[0, 1]$, będącą oceną jego jakości.

4 Wyniki eksperymentalne

Przeprowadziliśmy szereg eskperymentów zarówno na danych syntetycznych, jak i na rzeczywistych biologicznych danych białkowych. Aby ocenić przydatność stosowanych technik również w innych działach bioinformatyki, zbadaliśmy ich użyteczność do analizy sekwencji DNA.

W przeprowadzonych eksperymentach pokazujemy, że proponowane metody estymacji modelu uliniowienia zachowują się zgodnie z naszymi oczekiwaniami. W teście klasyfikacji przeprowadzonym z użyciem weryfikacji krzyżowej metoda rozkładów Beta okazała się być podobnej skuteczności, co estymator bayesowski używający mieszaniny rozkładów Dirichleta jako wiedzy *a priori*. Czas estymacji w naszej metodzie jest jednakże do 17 razy krótszy [18].

Zaproponowany model uwzględniający zależności oraz wiedzę *a priori* w postaci mieszaniny rozkładów Dirichleta, w teście klasyfikacji przeprowadzonym z użyciem weryfikacji krzyżowej okazał się być znacznie skuteczniejszy niż model bez zależności. Było to widoczne szczególnie w przypadku uliniowień zawierających wiele sekwencji (czyli macierzy uliniowienia zawierającej wiele wierszy) [18]. W przypadku liczby sekwencji mniejszej niż 400 (dla uliniowień białkowych) skuteczność estymatora bardzo spada. Jest to związane z faktem, że do wiarygodnej estymacji zależności między kolumnami potrzebne jest wiele przykładów w próbie losowej.

Przeprowadziliśmy także szereg eksperymentów polegających na poszukiwaniu fragmentów sekwencji testowych dobrze dopasowanych do badanych uliniowień. W eksperymentach tych pokazujemy wyższość Powtórnej Weryfikacji nad innymi metodami predykcji, również nad metodą p-wartość. Omawiane eksperymenty dotyczą sekwencji białkowych, jednakże jeden z nich został przeprowadzony na sekwencji DNA [19] i polegał na poszukiwaniu miejsc wiązania czynników transkrypcyjnych. Również w tym eksperymencie pokazana została wyższość metody Powtórnej Weryfikacji.

Pokazaliśmy również, jak użyć metody Oczekiwanej Powtórnej Weryfikacji do selekcji najlepszego z kilku możliwych modeli dla danego uliniowienia. W ten sposób powstaje model hybrydowy — dla każdego uliniowienia wybierany jest jeden (najlepszy) model, dla przykładu może być to model z zależnościami dla pewnych uliniowień, model PSSM stworzony metodą pseudozliczeń dla innych, dla jeszcze innych uliniowień model PSSM stworzony metodą rozkładów Beta. W przeprowadzonych eksperymentach pokazaliśmy, że predykcja fragmentów sekwencji testowych dobrze dopasowanych do badanych uliniowień z użyciem takiego modelu hybrydowego jest znacznie skuteczniejsza, niż dla modeli czystych oraz niż dla modeli hybrydowych uzyskanych innymi metodami.

W rozprawie odnosimy się również do ostatniego eskperymentu CASP (czyli CASP 6 [3] z 2004 roku). Pokazujemy, jak zachowuje się metoda lokalnych deskryptorów Fidelisa, jeśli wyposażyć ją w omawiane w rozprawie techniki analizy sekwencyjnej uliniowień białkowych. Porównujemy nasze wyniki z najlepszymi światowymi predykcjami nadesłanymi na CASP i stwierdzamy, że nasze predykcje struktur białkowych są dość dobrej jakości (w 8 na 23 przypadkach dostajemy predykcje lepsze niż serwer ROBETTA [6], uznawany za wiodący w tej dziedzinie).

5 Podsumowanie

Wyniki zaprezentowane w rozprawie mogą być z powodzeniem użyte jako część metody przewidywania struktury trójwymiarowej białek w oparciu o lokalne deskryptory. Dodatkowo pokazujemy ich użycie w innym dziale bioinformatyki, genomice regulatorowej, do wyszukiwania nowych miejsc wiązania czynników transkrypcyjnych.

W ramach prac powstały implementacje wszystkich omawianych w rozprawie metod. Ponieważ metoda lokalnych deskryptorów odniosła sukces w porównaniu z innymi metodami, które wzięły udział w ostatnim eksperymencie CASP, planujemy jej zgłoszenie do najnowszego eksperymentu CASP w 2006 roku.

Literatura

[1] PROTEINS: Structure, Function, and Genetics. Supplement 5, 2001. Articles published online in Wiley InterScience, 28 January 2002.

[2] PROTEINS: Structure, Function, and Genetics. Supplement 6: CASP 5, 2003.

[3] PROTEINS: Structure, Function, and Bioinformatics. Supplement 7: CASP 6, 2005.

[4] P. Agarwal and V. Bafna. Detecting non-adjoining correlations within signals in DNA. In *RECOMB'98*, pages 2–8, 1998.

[5] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. In *RECOMB'03*, pages 28–37, 2003.

[6] R. Bonneau, C. E. M. Strauss, C. A. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson, and D. Baker. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.*, 322(1):65–78, Sep 2002.

[7] M. P. Brown, R. Hughey, A. Krogh, I. S. Mian, K. Sjölander, and D. Haussler. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In L. Hunter, D. Searls, and J. Shavlik, editors, *ISMB-93*, pages 47–55, Menlo Park, CA, 1993. AAAI/MIT Press.

[8] M. L. Bulyk, P. L. F. Johnson, and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucl. Acids Res.*, 30(5):1255–1261, Mar 2002.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *J. Royal Statistical Soc. Series B*, 39:1–38, 1977.

[10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[11] C. Fraley and A. E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[12] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[13] N. I. Gershenzon, G. D. Stormo, and I. P. Ioshikhes. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res*, 33(7):2290–2301, 2005. Evaluation Studies.

[14] K. Karplus. Evaluating regularizers for estimating distributions of amino acids. *Proc Int Conf Intell Syst Mol Biol*, 3:188–196, 1995.

[15] U. Keich and P. A. Pevzner. Finding motifs in the twilight zone. In *RECOMB'02*, pages 195–203, 2002.

[16] O. D. King and F. P. Roth. A non-parametric model for transcription factor binding sites. *Nucl. Acids Res.*, 31(19):e116, Oct 2003. Evaluation Studies.

[17] A. Kryshtafovych and K. Fidelis. Local descriptors of protein structure. Part I. General approach and classification of local 3D regions in proteins. In preparation.

[18] S. Nowakowski, K. Fidelis, and J. Tiuryn. Introducing dependencies into alignments analysis and its use in protein local structure prediction. In R. Wyrzykowski, editor, *Proceedings of the Sixth International Conference on Parallel Processing and Applied Mathematics (PPAM'05), LNCS 3911, Springer*, pages 1106–1113, 2006.

[19] S. Nowakowski and J. Tiuryn. A new approach to assessing quality of predictions of transcription factor binding sites. To appear in Journal of Biomedical Informatics.

[20] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Computer Applications in Biosciences*, 12:327–345, 1996.

[21] R. L. Tatusov, S. F. Altschul, and E. V. Koonin. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *PNAS*, 91:12091–12095, 1994.