Autoreferat rozprawy doktorskiej pt.

Evolution of protein-protein interaction networks

Tytuł w języku polskim:

Ewolucja sieci oddziaływań białko-białko

Janusz Dutkowski

Tematem rozprawy jest analiza porównawcza sieci interakcji białko-białko pomiędzy gatunkami. Problem ten znajduje się w obszarze zainteresowań ewolucyjnej biologii systemów [1], dziedziny, której celem jest badanie ogólnych reguł organizacji i działania maszynerii komórkowej (w przeciwieństwie do szczegółowego badania jej elemetów). Nowe technologie eksperymentalne, między innymi system dwu-hybrydowy oraz spektrometria mas, pozwalają na badanie oddziaływań między białkami na skalę całego proteomu. W efekcie dla każdego zbadanego organizmu powstaje złożona sieć, w której węzłami są białka danego gatunku, a krawędziami - fizyczne interakcje między białkami. Sieci te wykazują ciekawe własności stanowiące bieżący temat teoretycznych badań biologów ewolucyjnych, a także matematyków i fizyków. Są one bezskalowe i charakteryzują się dużą modularnością oraz małą średnicą. Z praktycznego punktu widzenia, najważniejszym problemem jest systematyczna analiza dużych sieci. Wymaga ona rozwoju nowych modeli i narzędzi bioinformatycznych [2]. Istotną rolę pełnią tu metody genomiki porównawczej, zestawiające ze sobą dane różnych gatunków.

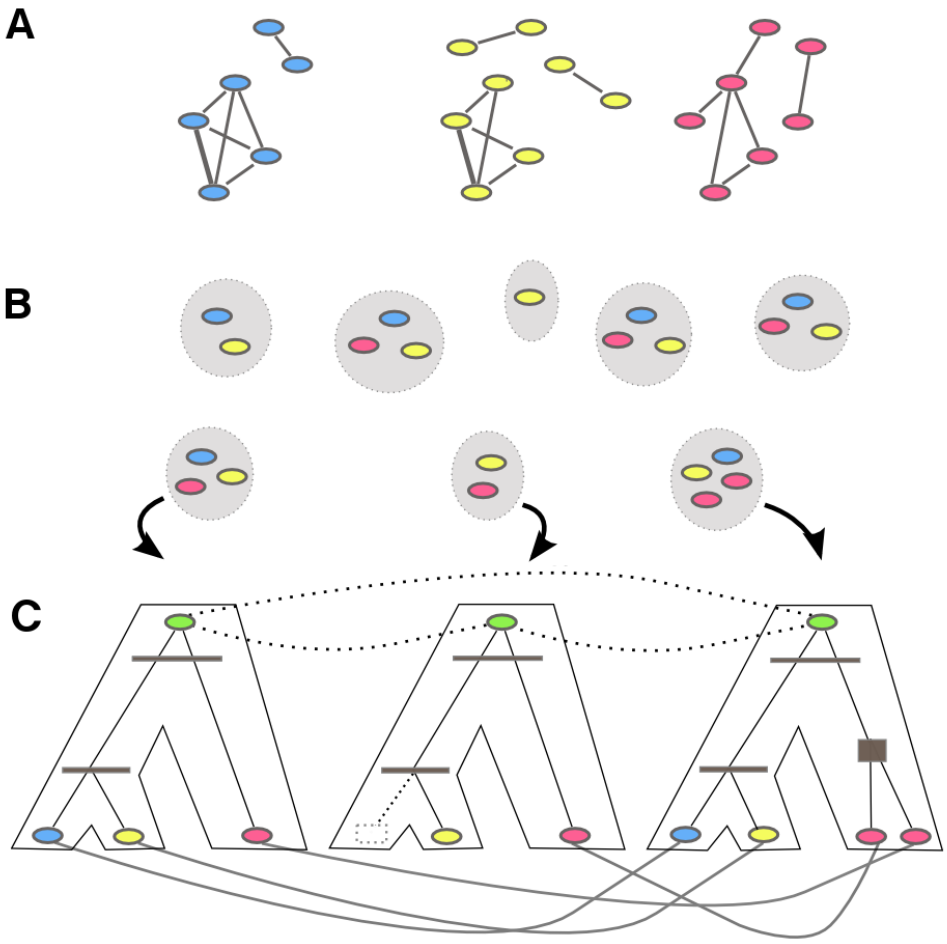
Pierwsze rozdziały rozprawy mają charakter wstępny. Przedstawiają metody analizy ewolucji sekwencji białek, wybrane modele losowych grafów oraz model sieci Bayesowskich. Kolejne trzy rozdziały opisują główny wynik badań, jakim jest opracowanie nowej metody porównywania sieci pomiędzy różnymi gatunkami, opartej na modelu ewolucyjnym i analizie filogenetycznej.

Metoda ta identyfikuje ewolucyjnie zachowane podsieci, które często odpowiadają funkcjonalnym kompleksom białkowych lub szlakom przekazywania sygnałów w komórce. W efekcie pozwala to na wyróżnienie w sieci współpracujących ze sobą modułów, odpowiedzialnych za kluczowe (silnie konserwowane) funkcje komórkowe. Osobnym zastosowaniem opracowanej metody jest przewidywanie interakcji w danym gatunku na podstawie interakcji u innych gatunków – uwzględniana jest przy tym relacja ewolucyjna pomiędzy odpowiednimi białkami.

Przedstawię teraz zarys badań nad modelowaniem i analizą sieci interakcji białek oraz streszczę wyniki opracowane w rozprawie.

Modele ewolucji sieci interakcji białek. Matematyczną reprezentacją sieci interakcji białek jest graf, w którym węzły odpowiadają białkom, a krawędzie reprezentują oddziaływania między nimi. Do opisu i porównywania sieci często wykorzystuje się globalne własności grafów, takie jak: rozkład stopni wierzchołków, średnica, a także tzw. współczynnik klasteryzacji. W kontekście zrozumienia procesów kształtowania się sieci interakcji białek, interesujące jest porównanie własności obserwowanych sieci z własnościami grafów losowych, przy ustalonym modelu stochastycznym. W odniesieniu do wielu rzeczywistych sieci białek (także wielu innych występujących w naturze sieci) zaobserwowano [3], że istotnie różnią się one od tych generowanych przez klasyczne modele grafów losowych (model Erdősa i Rényi oraz model Gilberta [4]). Powstała więc potrzeba opracowania nowych modeli, dokładniej opisujących obserwowane sieci. Jednym z lepiej zbadanych dotychczas modeli jest model sieci bezskalowych (o potęgowym rozkładzie stopni wierzchołków) zaproponowany przez Bollobása [5, 6, 7]. Model ten nie uwzględnia jednak biologicznych procesów zmieniających sieci białkowe.

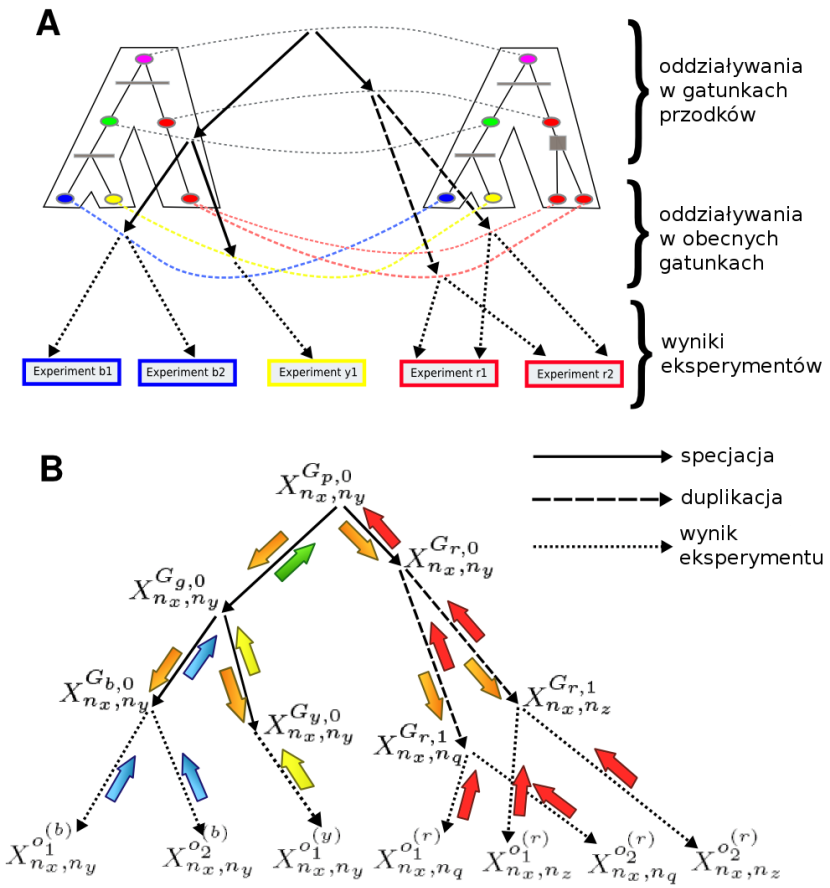
W modelowaniu sieci interakcji białek kluczowe jest uwzględnienie sposobu ich zmian w procesie ewolucji. Przyjmuje się, że obserwowane sieci powstały z mniejszych sieci przodków ewolucyjnych poprzez duplikacje białek oraz ich mutacje, zmieniające oddziaływania. Do opisu tego procesu zaproponowano tzw. model duplikacyjny [8, 9]. W rozprawie przedstawiona jest specyficzna wersja tego modelu [10], oparta dodatkowo na uzgodnionych drzewach filogenetycznych [11, 12]. Model ten jest zastosowany do wnioskowania o interakcjach między białkami na wcześniejszych etapach ewolucji.



Rysunek 1: Schemat algorytmu odtwarzania historii ewolucji sieci interakcji białek. W części (A) pokazane są przykładowe fragmenty wejściowych sieci trzech gatunków. W części (B) na podstawie porównania sekwencji białek nastepuje podział na rodziny. W części (C) dla każdej rodziny budowane jest drzewo filogenetyczne, które jest następnie uzgadniane z drzewem gatunków (grube drzewo). W wyniku uzgodnienia następuje ustalenie zjawisk duplikacji (pełny szary kwadrat), specjacji (linia pozioma) i strat białek (przerywany prostokąt). Na podstawie obserwowanych interakcji z dzisiejszych gatunków, drzew filogenetycznych oraz modelu ewolucji sieci, obliczone jest prawdopodobieństwo *a posteriori* krawędzi w sieci ewolucyjnego praprzodka (zielone węzły).

Porównywanie sieci pomiędzy gatunkami. Problem porównywania sieci biologicznych pomiędzy gatunkami znany jest w literaturze jako uliniowienie (ang. alignment) sieci. Celem lokalnego uliniowienia jest identyfikacja wspólnych, ewolucyjnie zachowanych podsieci (modułów), w których homologiczne (pochodzące od wspólnego przodka) białka mają podobne wzorce połączeń u różnych gatunków. Motywacją do tego typu poszukiwań jest powszechny pogląd, iż ewolucyjna konserwacja danego modułu sieciowego implikuje istotną funkcję tego modułu w systemie komórkowym. W jednej z pierwszych prac porównujących sieci białkowe [13], autorzy zaproponowali algorytm wyszukiwania zachowanych ewolucyjnie ścieżek prostych w sieciach dwóch gatuntów. Ścieżki interakcji odpowiadają tu biologicznym szlakom sygnałowym. Ta sama grupa badawcza zaproponowała algorytm odkrywania ewolucyjnie zachowanych, gęsto połączonych podsieci, potencjalnie odpowiadających kompleksom białkowym [14]. W artykule przeglądowym, podsumowującym rozwój metod uliniowienia sieci [15], zwrócono uwagę na istotne problemy związane z wcześniejszymi rozwiązaniami. Większość z nich mogła porównywać najwyżej dwie lub trzy sieci jednocześnie, ze względu na wykładniczy, względem liczby sieci, rozmiar sformułowanego zadania obliczeniowego. Żadna z metod nie była także, w istotny sposób, oparta na modelu ewolucji interakcji, a jedynie na heurystycznych funkcjach oceny uliniowienia.

W rozdziale piątym rozprawy przedstawiam nową metodę lokalnego uliniowienia sieci interakcji białek wielu organizmów, która wykorzystuje model ewolucyjny [10]. Metoda odtwarza interakcje między białkami na każdym etapie ewolucji, łącznie z siecią praprzodka wszystkich wejściowych gatunków. Rozwiązanie to pozwala na wgląd w procesy kształtowania się sieci oraz umożliwia stawianie hipotez co do przebiegu ewolucji sieci interakcji u różnych gatunków. Na początku rekonstruowana jest filogenetyczna historię ewolucji każdej z rodzin białek występujących w wejściowych gatunkach (por. Rys. 1). Na podstawie drzew filogenetycznych oraz modelu ewolocji budowana jest sieć Bayesowska (por. [16]), reprezentująca zależności pomiędzy interakcjami na poszczególnych etapach ewolucji (por. Rys. 2). Ze względu na drzewiastą strukturę otrzymanej sieci Bayesowskiej, możliwe jest zastosowanie efektywnego algorytmu Pearla [17] do obliczenia rozkładu *a posteriori* każdej zmiennej losowej, pod warunkiem ustalonych zmiennych (danych o obserwowanych interakcjach). W odtworzonej sieci praprzodka identyfikowane są najsilniejsze składowe, które odpowiadają ewolucyjnie zachowanym modułom. Opracowana metoda zastała zastosowana do uliniowienia sieci drożdża (*S. cerevisiae*), muchy (*D. melanogaster*) i robaka (*C. elegans*). Algorytm zi-



Rysunek 2: (A) Schemat ewolucji interakcji pomiędzy elementami dwóch rodzin białek. Ewolucja poszczególnych rodzin jest zadana poprzez drzewa filogenetyczne. Pojawienie się krawędzi (przerywane łuki) na danym etapie ewolucji zależy od istnienia interakcji na poprzednim etapie. Potencjalna interakcja pomiędzy białkami praprzodka (fioletowe węzły) zostanie zachowana, z pewnym prawdopodobieństwem, pomiędzy odpowiednimi białkami w sieci czerwonej lub zielonej (po specjacji). Podobnie dzieje się w przypadku duplikacji czerwonego białka. W obu przypadkach możliwe jest też pojawienie się nowej krawędzi. Każdą krawędź z obecnych dzisiaj gatunków obserwujemy, z pewnym prawdopodobieństwem, w danych eksperymentalnych. Z określonym prawdopodobieństwem pojawiają się tam także fałszywe krawędzie. (B) Każdej potencjalnej interakcji, oraz każdej danej eksperymentalnej przypisana jest zmienna losowa. W ten sposób powstaje sieć Bayesowska o strukturze drzewa. Przy użyciu algorytmu Pearla, który wysyła lokalne wiadomości między węzłami (kolorowe strzałki), liczone jest prawdopodobieństwo *a posteriori* każdej krawędzi pod warunkiem wszystkich danych ekperymentalnych.

dentyfikował wiele zachowanych modułów, spośród których część dobrze pasuje do znanych kompleksów drożdżowych (według referencyjnej bazy MIPS). Eksperymenty przedstawione w rozprawie wykazały również, że zaproponowana metoda wykrywa więcej potwierdzonych modułów niż wcześniejszy algorytm [14].

Integracja danych eksperymentalnych i przewidywanie nowych oddziaływań. W rozdziale szóstym przedstawiam rozwinięcie wcześniejszej metody i zastosowanie jej do integracji i poprawiania odziaływań białko-białko w badanych gatunkach [18]. Ze względu na ograniczenia technologiczne, aktualne dane eksperymentalne o interakcjach są niekompletne, a także zawierają dużo wyników fałszywych. Uwzględnienie danych z wielu esperymentów, a także ich integracja pomiędzy gatunkami pozwala ocenić wiarygodność poszczególnych interakcji oraz przewidzieć brakujące oddziaływania. Zaproponowany model w naturalny sposób umożliwia przepływ informacji o interakcjach pomiędzy parami białek, uwzględniając relacje ewolucyjne (por. Rys. 2). Przy jego użyciu obliczyłem zintegrowane sieci interakcji dla siedmiu organizmów eukariotycznych: rzodkiewnika (*A. thaliana*), drożdża (*S. cerevisiae*), muchy (*D. melanogaster*), robaka (*C. elegans*), szczura (*R. norvegicus*), myszy (*M. musculus*) oraz człowieka (*H. sapiens*). Ekperymenty przeprowadzone z użyciem referencyjnych baz danych (m.in. Gene Ontology, MIPS i HPRD) sugerują, że zintegrowane zbiory danych są lepsze niż wejściowe. Udało się także przewidzieć wiele interakcji w oparciu o dane z innych gatunków. Wyniki wykazują również przewagę proponowanej metody nad wcześniej zaproponowanymi podejściami [19, 20]. Rozprawa zawiera opis kilku przykładów znanych kompleksów białkowych, w obrębie których udało się skutecznie przewidzieć interakcje. Opierając się na wynikach eksperymentów, przedstawiam też hipotezę o domniemanych interakcjach w słabo poznanym kompleksie SWI/SNF (remodelującym chromatynę) w *A. thaliana*.

Estymacja parametrów modelu ewolucyjnego. W rozdziale siódmym przedstawiam algorytm do estymacji parametrów opracowanego modelu ewolucyjnego. Algorytm stosuje schemat Expectation Maximization (EM), służący do estymacji parametrów według zasady największej wiarygodności w sytuacji, gdy niektóre zmienne są ukryte [21]. Algorytm dla opisanego modelu ewolucji sieci jest rozwinięciem dobrze znanego algorytmu Bauma-Welcha, który estymuje parametry w ukrytych modelach Markowa. Opracowany al-

gorytm jest zastosowany do estymacji parametrów dwóch modeli ewolucji interakcji. Model pierwszy zakłada, że interakcje pomiędzy elementami danych rodzin są zachowane ewolucyjnie. Model drugi odpowiada losowej ewolucji sieci. Udało się pokazać, że parametry ewolucji obliczone dla znanych szlaków sygnałowych i kompleksów białkowych różnią się istotnie od tych, obliczonych na podstawie losowo wybranych fragmentów sieci. Jest to ciekawy wynik, który może posłużyć do konstrukcji metody dyskryminacyjnej przypisującej rozpatrywaną parę rodzin do jednego z modeli według zasady największej wiarygodności. Na tej podstawie budowana jest sieć współpracujących ze sobą (w sensie interakcji) rodzin białek. Porównanie z bazą Gene Ontology wykazało, że moduły w tej sieci mają istotne statystycznie anotacje funkcjonalne.

Ostatni rozdział rozprawy zawiera podsumowanie przeprowadzonych badań i głównych wyników. Przedstawia też możliwe kierunki rozwoju opracowanej metody oraz jej dalszych zastosowań.

Literatura

[1] Mónica Medina, (2005). Colloquium Paper: Systematics and the Origin of Species: Genomes, phylogeny, and evolutionary systems biology. Proc Natl Acad Sci U S A. 102(Suppl 1): 6630-6635.

[2] Panchenko, A., Przytycka, T., Eds. (2008). Protein-protein Interactions and Networks: Identification, Computer Analysis, and Prediction. Springer-Verlag London.

[3] Yook, S.H., Oltvai, Z.N., and Barabasi, A.L. (2004). Functional and topological characterization of protein interaction networks. Proteomics 4, 928-942.

[4] B. Bollobás, B. (2001). Random Graphs, Second Edition. Cambridge University Press.

[5] Bollobás, B., Riordan, O. (2002). Mathematical results on scale-free random graphs. In Handbook of Graphs and Networks. Wiley-VCH, Berlin.

[6] Bollobás, B., O. Riordan, O. (2004). The diameter of a scale-free random graph. Combinatorica, 24(1):5-34.

[7] Bollobás, B., Riodan, O, Spencer, J., Tusnady, G. (2001). The degree sequence of a scale-free random graph process. Random Structures and Algorithms 18:279-290.

[8] Sole, R. V., Pastor-Satorras, R., Smith, E., and Kepler, T. B. (2002). A model of large-scale proteome evolution. Advances in Complex Systems, 5, 43.

[9] Pastor-Satorras, R., Smith, E., and Sol, R. V. (2003). Evolving protein interaction networks through gene duplication. J Theor Biol, 222(2).

[10] Dutkowski, J., Tiuryn, J., (2007). Identification of functional modules from conserved ancestral protein-protein interactions. Bioinformatics (Oxford, England) 23(13):i149-58.

[11] Page, R., Charleston, M. (1997). From gene to organismal phylogeny: reconciled trees and th gene treespecies tree problem. Mol. Phylogenet. Evol., 7 (2):231-40.

[12] Górecki, P., Tiuryn, J. (2006). DLS-trees: a model of evolutionary scenarios. Theor. Comput. Sci., 359, 378-399.

[13] Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R., and Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A, 100(20), 11394-11399.

[14] Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M., and Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A, 102(6):1974 - 1985.

[15] Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. Nature Biotechnology, 24(4):427-433.

[16] Neapolitan, R.E. (2003). Learning Bayesian Networks. Prentice Hall.

[17] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Netowrks of Plausible Inference. Morgan Kaufmann.

[18] Dutkowski, J., Tiuryn, J. (2009). Phylogeny-guided interaction mapping in seven eukaryotes. (zgłoszona do publikacji)

[19] Michaut, M. and Kerrien, S. and Montecchi-Palazzi, L. and Chauvat, F. and Cassier-Chauvat, C. and Aude, J. C. and Legrain, P. (2008). InteroPorc: Automated Inference of Highly Conserved Protein Interaction Networks. Bioinformatics (Oxford, England), 24(14):1625-1631.

[20] Liu, Y., Liu, N., Zhao, H. (2005). Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics (Oxford, England), 21(15):3279-3285.

[21] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39, 1-38.