PROTEIN SEDONDARY STRUCTURE PREDICTION BASED ON NEURAL NETWORK

Karel Sedlář

Master Degree Programme (1), FEEC BUT

E-mail: xsedla74@stud.feec.vutbr.cz

Supervised by: Helena Škutková

E-mail: skutkova@feec.vutbr.cz

Abstract: Detection of protein's secondary structure is an essential process in the reconstruction of their 3D models, which leads to better understanding of their chemical and physical properties. Due to this information we are able to reveal their biological role. Unfortunately, methods for accurate determination of the secondary structure as X-ray crystallography and NMR are very expensive and time consuming. This makes them impossible to be used in many cases. Mathematical models are an alternative for estimating the structure. Neural network appears to be an effective tool for this operation.

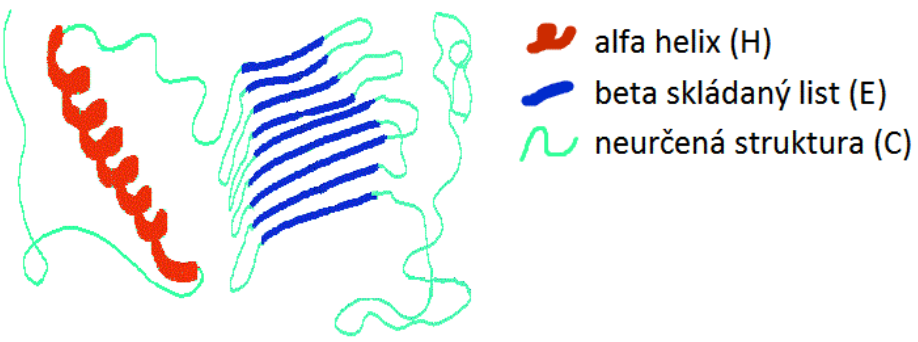
Keywords: protein, secondary structure prediction, neural network

1. ÚVOD

Studium proteinů je neodmyslitelnou součástí molekulární biologie. Stanovení prostorové struktury, neboli konformace, pak hraje důležitou roli v pochopení jejich chemických a fyzikálních vlastností. Na základě těchto vlastností lze proteiny klasifikovat a zjišťovat jejich biologickou funkci. Pro určení výsledné třírozměrné struktury je nutné znát motivy nejnižšího stupně 3D struktury, tedy struktury sekundární. Přesné stanovení je možné pouze pomocí rentgenové krystalografie nebo nukleární magnetické resonance, což jsou metody finančně i časově velice náročné. Vhodnou alternativu však mohou poskytnout odhady sekundární struktury výpočetními nástroji bioinformatiky. Jelikož není známý přesný matematický popis, jakým dochází k vytváření sekundární struktury na základě struktury primární tj. sekvence aminokyselin, lze s výhodou využít neuronové sítě, které jsou samy schopné naučit se ideální vzorec predikce na základě předložené učební množiny. Ta je v tomto případě tvořena proteiny, jejichž sekundární struktura je již známa.

2. STRUKTURA PROTEINU

Pro účely predikce se uvažují dva základní motivy a to α-helix (helix) a β-skládaný list (strand). Ostatní motivy jsou označeny za neurčené (coil).[1]



Obrázek 1: Motivy ve struktuře proteinů

2.1. ZÁPIS PRIMÁRNÍ STRUKTURY

K zápisu primární struktury proteinu se používá posloupnost jednopísmenových IUPAC kódů pro aminokyseliny. Ty jsou zapsány dle konvence ve směru od N-konce k C-konci proteinu. Tyto znaky mají číselný ekvivalent, který se pro účely neuronových sítí použije na binarizaci zápisu všech 20 biogenních aminokyselin.

Tabulka 1: Příklady zápisu aminokyselin

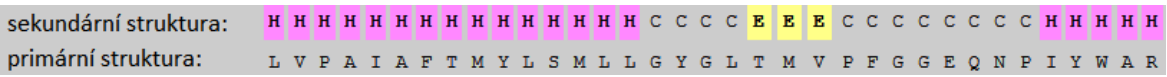
Aminokyselina	IUPAC kód	Číselný kód	Binární kód
Alanin	A	1	10000000000000000000
Arginin	R	2	01000000000000000000
Cystein	C	5	00001000000000000000

2.2. ZÁPIS SEKUNDÁRNÍ STRUKTURY

Motivy sekundární struktury se zapisují obdobně, také sekvencí jednopísmenových DSSP kódů, kde každý jeden znak vyjadřuje, jakou strukturu aminokyselina tvoří. Pro účely neuronové sítě je nutné tento zápis také binarizovat.

Tabulka 2: Zápis sekundární struktury

Motiv	Obecné označení	DSSP kód	Binární kód
Neurčeno	Coil	C	100
β-skládaný list	Strand	E	010
α-helix	Helix	H	001



Obrázek 2: Ukázka zápisu primární a sekundární struktury proteinu

3. NEURONOVÁ SÍŤ

Neuronové sítě jsou algoritmy složené z jednotlivých prvků - neuronů, jejichž matematický popis existuje již od 40. let minulého století. Pokud jim předložíme učební množinu proteinů s již známou strukturou jsou schopny vhodně aproximovat systém predikce sekundární struktury.[2]

3.1. UČEBNÍ SOUBOR DAT

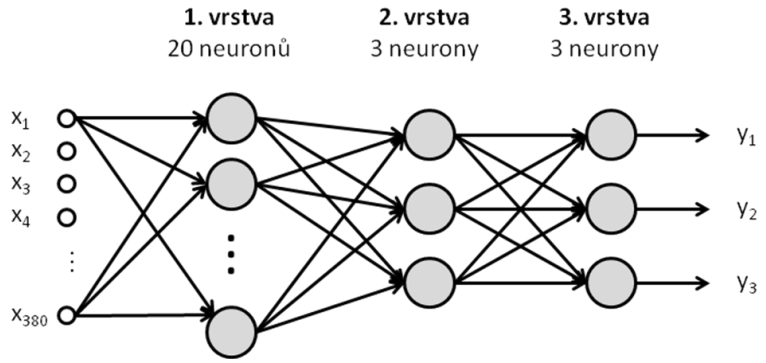
Ačkoliv bylo publikováno několik algoritmů, které udávají přesnost predikce přesahující 70%, v reálném uplatnění takové přesnosti nedosahují, protože byly sestrojeny na malém učebním souboru RS126 [1], kde ověřování probíhalo pouze na několika málo proteinech. Bylo dokázáno, že při použití novějšího a většího souboru CB513 klesá schopnost správné predikce u těchto sítí k 50% [3]. Jiné algoritmy fungují online, přičemž učení probíhá na základě velmi podobných proteinů poskytnutých BLAST. Jejich přesnost dosahuje i 80% [4], avšak pokud nejsou dostupné homologní proteiny se známou strukturou i jejich přesnost klesá pod 60%. Jako učební soubor byl tedy zvolen CB513 obsahující 513 vzájemně evolučně vzdálených proteinů jejichž sekundární struktura byla zjištěna rentgenovou krystalografií či NMR.

3.2. STRUKTURA NAVRŽENÉ SÍTĚ

Sekundární struktura není podmíněna pouze jednou aminokyselinou, ale uplatňuje se i vliv sousedních aminokyselin. Optimálních výsledků bylo dosaženo s oknem délky 19, tedy motiv sekundární struktury pro danou aminokyselinu je odhadnut na základě 9 předcházející a 9 následujících aminokyselin. Protože binární zápis jedné aminokyseliny je tvořen vektorem 20 hodnot, celkový počet vstupů je tak 19·20=380. Výstupem musí být vektor 3 hodnot (viz. Tabulka2), tedy výstupní vrstva obsahuje 3 neurony. Mezi ni a vstup jsou vnořeny ještě dvě vrstvy, první o 20, druhá o 3 neuronech. Celkově jsme tak definovali třívrstvou perceptonovou dopřednou síť se zpětným šířením chyby ve fázi učení.

Na množinu vstupů x, připadá množina výstupů y:

$$\vec{x} = [x_1, x_2, \dots, x_{380}]^T \qquad \vec{y} = [y_1, y_2, y_3]^T \tag{1}$$



Obrázek 3: Motivy ve struktuře proteinů

Pro učení vícevrstvých dopředných sítí je nezbytné, aby přenosové funkce měli spojitou derivaci. Jako vhodné se tudíž jeví použití různých typů sigmoidní funkce. Při konstrukci byla zvolena jako výchozí dvouvrstvá síť, která je pro predikci sekundární struktury používána nejčastěji. Přitom nejlepších výsledků bylo dosaženo, když charakteristika první vrstvy byla sigmoida a druhé vrstvy pak její logaritmovaná verze. Výsledky však obsahovaly podstatnou část špatně určené struktury na hranicích mezi množinami výstupů (C, E, H). Síť proto byla doplněna ještě o jednu vrstvu neuronů s přenosovou funkcí sigmoidy, která výrazně zlepšila pravděpodobnost správné predikce.

3.3. VÝSLEDKY

Učení a testování proběhlo podle zásad definovaných v [3]. Tedy učení proběhlo na 4/5 použitého souboru CB513 a testování na zbývajících 100 proteinech, které nebyly použity pro učení sítě. Učební cyklus byl zastaven po 213 epochách. Přičemž bylo dosaženo výsledné pravděpodobnosti správné predikce 66,7%.

4. ZÁVĚR

V dnešní době je dostupných několik algoritmů na predikci sekundární struktury proteinů, jejichž přesnost dosahuje i 80%, ovšem pouze pro proteiny jejichž blízce příbuzné homology již mají experimentálně zjištěnou strukturu. Stále však existuje mnoho proteinů, pro které není možné najít blízký protein s definovanou sekundární strukturou. Pro predikci u takovýchto proteinů se jeví jako nejlepší řešení použití neuronové sítě, jejíž učení probíhá na souboru vzájemně evolučně vzdálených proteinů. Výše definovaná síť se od běžně používaných liší třemi vrstvami, oproti běžným dvěma. Na použitém souboru proteinů vykazuje predikci s přesností téměř o 4% vyšší než síť definovaná podle stejných kritérii v [3].

REFERENCE

[1] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J *Mol Biology*, 1993; 232: 584-99.

[2] Holley, L.H., and Karplus, M. (1988). Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA. 86, 152-156.

[3] Avdagic Z, Purisevic E, Buza E, Coralic Z. Neural Network Algorithm for Prediction of Secondary Protein Structure. Acta Inform Med. 2009; 17(2): 67-70

[4] Buchan, D. W. A., Ward, S. M., Lobley, A. E., Nugent, T. C. O., Bryson, K., & Jones, D. T. (2010). Protein annotation and modelling servers at University College London. *Nucleic Acids Research*, *38*(Web Server issue), W563-W568. Oxford University Press.