PROTEIN HOTSPOT PREDICTION USING S-TRANSFORM - THRESHOLDING METHOD INFLUENCE

Jan Kašpárek
Doctoral Degree Programme (1), FEEC BUT
E-mail: xkaspa26@stud.feec.vutbr.cz

Supervised by: Ewaryst Tkacz
E-mail: tkacz@feec.vutbr.cz

Abstract: Protein hotspot localization is the first step on a path to fully control protein-protein interactions and by means of this directly influence course of various metabolic paths. In this paper we propose new thresholding approach that we call peak detection that could me implemented into other already existing algorithms. Tests results of our tests show that by using peak detection we succesfully add another degree of freedom while changing the algorithm efficiency only negligibly.

Keywords: Hotspot localization, proteins, signal processing, thresholding, S-transform

1 INTRODUCTION

Basically all metabolic paths in any organism rely on catalytic power of proteins at some point of their course. Only rarely these proteins are actually composed of a single protein unit. More often we can find these molecules to be oligomers composed of several monomer units. Number of these units can vary from two to tens of them [1].

Protein–protein interaction is therefore essential for their function. Let us have a closer look on how such an interaction actually takes place. Interaction among individual protein molecules is carried out by aligning certain areas of their surface. For obvious reasons these areas are called binding sites. They can be rather large and contain tens or even hundreds of residues. Although once we look even closer, we can find out that most off these residues do not exhibit any direct participation to the originating bond between protein units. In fact only few residues are responsible for the majority of binding free energy. These few residues are called hotspots – the name is based on their abundance of energy – and since their importance is self-evident they are of particular interest to scientific community.

However the biggest nuisance of hotspots is the fact that there is still no feasible way of their localization. At first scientist were limited to conducting strenuous wetwork experiments and developed a method known as Alanine Scanning Mutagenesis (ASM). This method is still considered to be the most reliable one and is often used as a reference. On the other hand it is also extremely financially demanding and time consuming. Details about ASM methodology can be found in [2]. Drawbacks of laboratory techniques represented by ASM and a few of its modifications created a niche for computational hotspot prediction techniques and since then this is still the direction in which hotspot prediction is evolving.

2 PROTEIN HOTSPOT PREDICTION ALGORITHM

In this paper we investigate possible benefits of a specific thresholding method here described as peak detection incorporated into hotspot prediction algorithm. The algorithm itself is based on signal processing techniques, namely S-transform, which is a powerful time-frequency analysis tool. Unlike

large number of existing algorithms, ours does not require any information about protein structure. All that is necessary is sequence of the examined protein and for reasons described later also sequences of several similar proteins.

2.1 NUMERICAL REPRESENTATION OF PROTEIN SEQUENCE

First, we need to convert sequence of residues to numerical signal. To do this we exploit a physical quantity called Electron-ion interaction potential (EIIP). This quantity describes average valence electrons' energy for each residue. However any measure of energy can only be non-negative. Sequence of purely non-negative values would introduce an undesirable DC bias into our data. Such a bias would manifest in a significant peak in signal's Fourier spectrum. Description of the next step will make it clear why this is unfavourable. The easiest way to suppress DC bias is to subtract signal's average value. This is also the way we deal with this problem ourselves.

2.2 CHARACTERISTIC FREQUENCY

According to Resonant Recognition Model (RRM) introduced in [3], proteins sharing common function also share a common frequency component in their Fourier spectra. This common frequency component is called characteristic frequency. However as an analogy to hotspots themselves characteristic frequency value is not an easy thing to determine.

This is where the need for similar proteins' sequences arises from. To calculate characteristic frequency value we multiply Fourier spectrum of examined protein with Fourier spectra of proteins sharing a common function with it. For sake of simplicity we call these proteins related proteins. Described calculation can be expressed in form of equation [4]:

$$S(e^{j\omega}) = |X_1(e^{j\omega}) \cdot X_2(e^{j\omega}) \cdot \ldots \cdot X_1(e^{j\omega})| \tag{1}$$

where $X_i(e^{j\omega})$ is Fourier spectrum of i-th related protein and $S(e^{j\omega})$ is consensual spectrum. To ensure equal length of all protein sequences we pad the shorter ones with zeroes.

However there is no way to tell how many related proteins' sequences reliable result will require. We need to keep multiplying the spectra until there is only one significant peak left. This peak then lies on characteristic frequency. It can happen that all you need is just one related protein, on the other hand it can even be more then ten. Furthermore numbers are not everything here. Proper selection of individual related proteins is even more important. By selecting incorrectly, you can still get a single peak in spectrum, but it will almost definitely be on different frequency. In general there is no definite way to tell how to choose related proteins properly. We used proteins of parallel function from different organisms. So when we were examining human growth hormone receptor, we used bull growth hormone receptor, chicken growth hormone receptor and others like this as its related proteins.

2.3 S-TRANSFORM

S-transform is a time-frequency analysis tool with an easy-to-read output similar to Short Time Fourier Transform (STFT) and at the same time retains some properties of wavelet transform like frequency dependent resolution.

Mathematically, ST is defined as follows [5]:

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t)\omega(\tau - t, f)e^{-j2\pi ft}dt \tag{2}$$

where $x(t)$ is time varying signal as the transformation's input, $S(\tau, f)$ is sample of the output corresponding to time shift τ and frequency f. Window ω is then defined as

$$\omega(t,\sigma) = \frac{1}{\sigma(f)\sqrt{2\pi}} e^{\frac{t^2}{2\sigma^2(f)}} \tag{3}$$

where σ is window width defined as

$$\sigma(f) = \frac{1}{|f|}. \tag{4}$$

Fig. 1 shows typical S-transform output of protein sequence. This data have been computed from human growth hormone receptor. We can see that most of the performance occurs on higher frequencies, that is an attribute common among proteins in general. The final signal we use for hotspot prediction is a cross-section along the characteristic frequency. Example of its waveform is shown later on in Fig. 2.

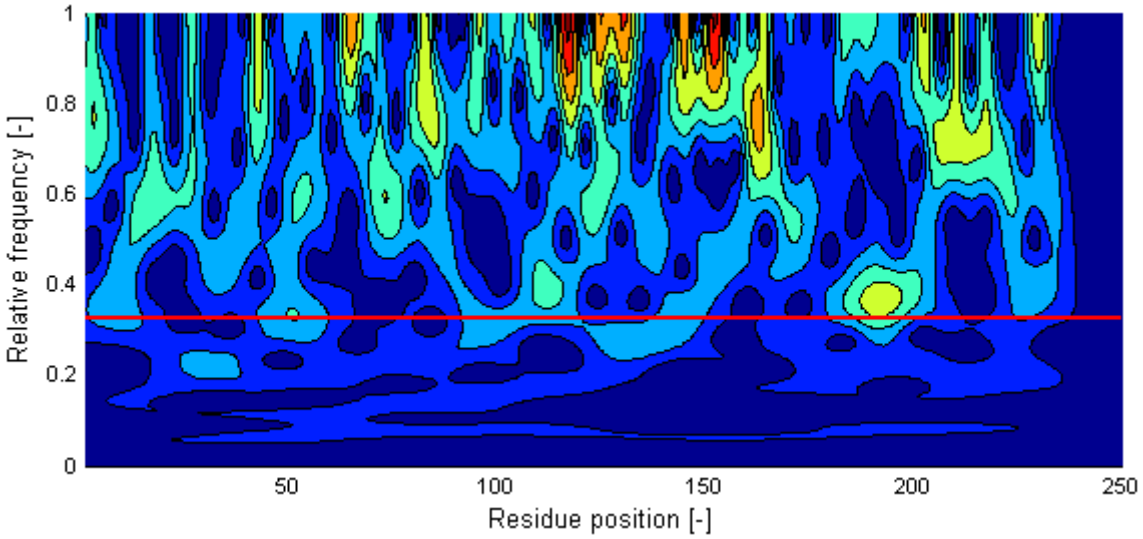

Figure 1: S-transform output of human growth hormone. Red line denotes characteristic frequency value determined in previous step. In this case 0.3231. Area of zero performance due to zero padding is not displayed.

3 FINAL HOTSPOT LOCALIZATION APPROACHES

So far we have successfully acquired a signal that will represent a criterion for final hotspot localization. Here we propose two ways of actual localization. Both will be described in following sections.

3.1 THRESHOLDING

An obvious method is simple thresholding. Since the cross-section of ST spectrum describes a measure of performance provided by individual residues, it is logical that energetically rich residues like hotspots should boast rather high performance. We therefore declare threshold value equal to the average performance and any above-threshold residue is then deemed a hotspot. We use this method as a baseline.

3.2 PEAK DETECTION

The idea of this approach is based on the fact, that hotspot energy richness might apply only in comparison with other closely positioned residues. In other words hotspot can be energetically rich in comparison to its neighbours but there might still be even richer residues somewhere else in the sequence that are not hotspots themselves.

To take this possibility into consideration we detect peaks in the waveform shown in Fig. 2. More precisely we only detect above-threshold peaks and a certain number of their neighbouring residues. So we still use a form of thresholding in this approach. However this time even residues under the threshold might be deemed hotspots if they are close enough to an above-threshold peak. Again the threshold value is equal to average performance to filter out smallest of peaks.

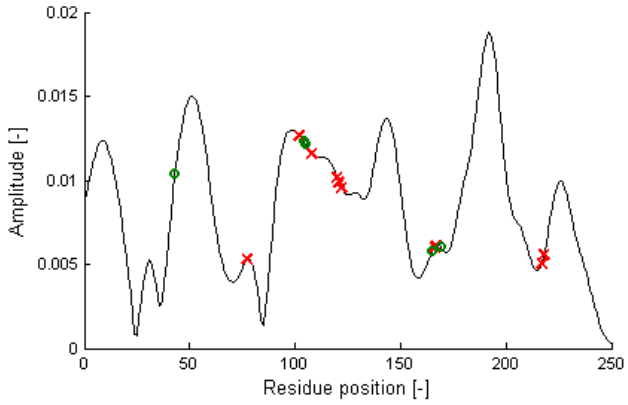

Figure 2: Slice of ST spectrum at characteristic frequency. Red crosses (10) denote a priori known non-hotspot residues, while green circles (5) denote a priori known hotspot residues. Area of zero performance due to zero padding is not displayed.

4 RESULTS

The thresholding approaches were tested on the dataset composed by Nguyen et al. [6] as a learning set for their own hotspot localization approach. It was put together with an aim to present a representative sample across various protein families. However we had to restrict ourselves only to proteins for which we were able to determine characteristic frequency properly. Our final dataset is therefore composed of twelve distinct proteins.

4.1 INFLUENCE OF NEIGHBOURHOOD WIDTH

Table 1 sums up results achieved with various widths of peak neighbourhood. Here, neighbourhood of width zero is only the peak residue itself while neighbourhood of width one adds one residue on either side. The rest continues in similar fashion.

We can see that with growing neighbourhood both sensitivity and positive predictive value (PPV) grow as well. However PPV reaches its maximum at width 10 (11) and since then keeps dropping slowly while Sensitivity stops its rise at width 13 and remains stable from there on.

4.2 THRESHOLDING VERSUS PEAK DETECTION

Simple thresholding does not operate with any parameters beside actual threshold value. This is set at the average performance, same goes for peak detection. With this approach we achieve sensitivity of

Neigh. width [residues]	Sensitivity [%]	PPV [%]	Neigh. width [residues]	Sensitivity [%]	PPV [%]
0	16.67	20.83	8	75.50	46.65
1	24.94	28.47	9	83.63	46.23
2	33.47	39.58	10	85.71	46.78
3	49.03	45.61	11	85.71	46.78
4	49.03	38.19	12	86.90	46.64
5	56.51	37.86	13	88.10	46.18
6	65.42	40.51	14	88.10	45.57
7	69.86	38.54	15	88.10	43.90

Table 1: Results obtained for different neighbourhood widths.

85.32 % and PPV of 49.7 %. This means that peak detection can achieve higher sensitivity, however it falls behind in terms of PPV.

5 CONCLUSION

We have applied two distinct hotspot localization approaches based on different point of view of hotspots' energy richness. Even though the results of both are very similar and peak detection achieves slightly higher sensitivity, it is PPV that is favoured in scientific community. Therefore we conclude that peak detection does not bring any significant benefits, however it adds another degree of freedom which might be useful option of fine-tuning particular hotspot detection algorithms. Still it would be interesting to apply peak detection to different signals intended for hotspot localization and see the results there.

ACKNOWLEDGEMENT

This paper has been supported by the Program Project UMO-2012/07/B/ST6/01238 and authors would like to express their appreciation forthat.

REFERENCES

[1] Rajamani, D., Thiel, S., Vajda, S., Camacho, C.J.: Anchor residues in protein-protein interactions. Proc. Natl. Acad. Sci. U. S. A. 101, 11287–92 (2004).

[2] Bogan, A.A., Thorn, K.S.: Anatomy of hot spots in protein interfaces. J. Mol. Biol. 280, 1–9 (1998).

[3] Veljković, V., Ćosić, I., Dimitrijević, B., Lalović, D.: Is It Possible to Analyze DNA and Protein Sequences by the Methods of Digital Signal Processing? Trans. Biomed. Eng. BME-32, 337–341 (1985).

[4] Ramachandran, P., Antoniou, A.: Identification of Hot-Spot Locations in Proteins Using Digital Filters. J. Sel. Top. Signal Process. 2, 378–389 (2008).

[5] Stockwell, R.G., Mansinha, L., Lowe, R.P.: Localization of the Complex Spectrum: The S Transform. Trans. Signal Process. 44, 998–1001 (1996).

[6] Nguyen, Q., Fablet, R., Pastor, D.: Protein Interaction Hotspot Identification Using Sequence-Based Frequency-Derived Features. Trans. Biomed. Eng. 60, 2993–3002 (2013).