What does the resonant recognition model tells us about Myosin Binding Protein C?

Reyder Risco Soria, Elena Pirogova, Jose Luis Hernandez Caceres and Irena Cosic.

CECAM-ISCMH, Havana, Cuba

RMIT, Melbourne, Australia.

Abstract

Objetivos: comparar las predicciones teóricas con datos conocidos acerca de la proteína C enlazadora de Miosina (Myosin Binding Protein C, MyBPC). **Metodos**: Se aplicó un análisis por modelo de reconocimiento resonante (resonant recognition model, RRM) para estudiar un conjunto de secuencias MyBPC con el objetivo de identificar las frecuencias relevantes. Un análisis de sitios calientes con análisis de Fourier de "grano grueso" se aplicó para predecir las regiones que mayormente contribuyen a la función. **Resultados**: Se encontraron dos frecuencias características: f= 0.226, aparentemente asociada a la isoforma cardiaca de la proteína, y f= 0.3667, que corresponde a MyBPC de modo inespecífico. Las regiones críticas en la proteína que fueron predichas por el método incluyeron la región de fosforilación, así como otras áreas cercanas o que incluyen muchas de las mutaciones que provocan cardiopatía familiar. Una secuencia "artificial" con todas las mutaciones puntuales descritas reportadas en la literatura mostró un pico mucho menor en el pico f=0.226. **Conclusiones**: al menos existen dos frecuencias de RRM asociadas a la función de la MyBPC, siendo f=0.226 la mas específica para la función cardiaca. Aproximadamente el 15% de las mutaciones puntuales descritas que aparentemente conducen a cardiopatía hipertrófica, una inserción especifica de la isoforma cardiaca, dos sitios de fosforilación y una mutación que conduce a la desnaturalización irreversible del dominio C5, caen dentro de las regiones de mayor contribución a la frecuencia f=0.226. Los autores sugieren que un enfoque basado en RRM seria útil para explicar o predecir otras enfermedades asociadas a mutaciones.

Abstract

Aim: to compare theoretical predictions with known data about Myosin Binding Protein C (MBPC). **Methods**: resonant recognition model (RRM) analysis was performed on MBPC from different types in order to identify relevant frequencies. Coarse Grained hot spot Fourier analysis was used for predicting regions of the sequence with major contribution to the function. **Results**: Two characteristic frequencies were found: f= 0.226, apparently associated to the cardiac isoform and f= 0.3667, non-specifically correspondent to MBPC. Critical areas in the sequence were predicted for the region of phosphorylation as well as for areas where several of the most important disease-causing mutations have been described. An artificial "sequence" with all point mutations reported in literature leads to a significant reduction of the peak at f=0.226. **Conclusions**: at least two RRM frequencies are associated to MYBPC function, being f=0.226 the most specific for cardiac function.

Nearly 15% of the point mutations purportedly associated to CMH4, a cardiac specific insert, two phosphorylation sites and a mutation leading to an irreversible denaturation of the dominion C5 fall into the regions of major contribution to f=0.226. Authors suggest that an RRM based approach could be useful for explaining or predicting other mutation-associated diseases.

Introduction

Myosin binding protein-C (MYBPC) is an accessory protein from the thick filament that is present in almost every type of vertebrate muscle. MYBPC function has not been yet clarified, but strongly evidence a point to a regulatory component in muscle contractility [1].

It has been shown that cardiac MYBPC is a target site to phosphorylation as response to different inotropic influences, including sympathetic stimuli that induce cardiac MYBPC phosphorylation through cAMP dependent protein kinase (PK). At the same time, there is a group of mutations in the gene for the cardiac isoform of MYBPC associated to Type 4 Familial Hypertrophic Cardiomyopathy (CMH4), a disease related to mutations in proteins participating in cardiac contractile machinery. Familial Hyperthrophic Cardiopathy is the most common cause of sudden cardiac death in young athletes. Its prevalence is believed to be about 0.2%, or 1 in 500[1-4].

However, in spite of the abundant evidence suggesting its importance for cardiac function, little is known about MYBPC function. Thus, several studies deal with effects of PK on cardiac contractility, but the role of cardiac MYBPC as part of the response to PK activation has not yet been elucidated [1-4].

Likewise, the mechanisms leading from MYBPC mutations into cardiac function impairment have not been clarified as yet. Presence of protein truncating mutations has suggested possible expression of truncated proteins in patient's muscle. The case of point mutations, deletions or insertions has remained unsolved hitherto. The precise mechanism by which many of these mutations cause FHC remains unsolved, but in some cases the degree to which these mutations may affect the structure and function of MYBPC can be inferred from sequence comparisons [4].

In any case, it is worth of notice the relatively large number of point mutations leading to disease.

One of the advantages of some bioinformatics methods lies on the possibility to make inferences about a protein, even when little is known about its function.

In this study we have employed the Resonant Recognition Model (RRM) approach to the analysis of MYBPC protein sequences. The method is based on the idea that characteristic frequencies must be shared by proteins with common function or interaction. In comparison with other bioinformatics methods used for computational

studies of protein structure-function relationships, the RRM brings up a number of advantages, such as [5-10]:

- The possibility to support for the presence (or absence) of a common function within a set of proteins.
- The possibility to theoretically explore the consequences of a certain mutation for the function of a protein.
- The possibility to find out the amino acids in a protein that are crucial for its function.

We have identified MYBPC characteristic frequencies corresponding to their common functions, predicted the positions of the critical regions in the amino-acidic sequence that are relevant to these functions and contrasted them with known facts from literature. We regard that the obtained results by means of the RRM analysis are in conformity with current knowledge about MYBPC obtained from both experimental and theoretical evidences. Furthermore, our results provide additional information about specific features of cardiac and muscle MYBPC isoforms.

We hypothesize that the application of the RRM may shed some light into the understanding of the functional behavior of MYBPC and its many modifications leading, or not, to pathological conditions. The fact that its implementation is computationally affordable, suggest that this approach may be extended for further clinical and research purposes.

Materials and Methods

The crucial problem of understanding the rules of encrypting of a protein bioactivity within the amino acid sequence still remains an unsolved goal. Thus, there is an urgent need for robust theoretical approaches for computational analysis of structural and functional performance of different protein families.

The RRM is a physic-mathematical approach that interprets protein sequence's linear information using digital signal processing [11-14]. The RRM suggests that protein interactions represent the transfer of the resonant energy between the interacting molecules at the frequency specific for each observed function/interaction. In the RRM the protein primary structure is represented as a numerical series by assigning to each amino acid in the sequence a physical parameter value relevant to the protein's biological activity. The RRM hypothesized that there is a significant correlation between spectra of the numerical presentation of amino acids and their biological activity [13, 14]. Though, it has been found in extensive researches, that protein with the same biological function have a common frequency in their numerical spectra. This frequency was found then to be a characteristic feature for protein biological function or interaction.

The RRM procedure involves two stages of calculations. Firstly, the original amino acid sequence is transformed into the numerical sequence by assigning to each amino acid a particular value of the physical parameter relevant to the protein biological function. The assignment of a particular number for each amino acid in the molecule is a crucial step in all these calculations. In this study, the energy of delocalized electrons (calculated as the electron-ion interaction pseudo-potential, EIIP) of each amino acid residue is used. Thus, the resulting numerical series represents the distribution of the free electrons energies along the protein. At the second stage, these numerical series are analyzed by digital signal analysis methods, in particular by Discrete Fourier Transform (DFT), in order to extract information pertinent to the protein biological function. A multiple cross-spectral function is defined and calculated to obtain the common frequency components from the spectra of a group of proteins. Peaks in such function denote common frequency components for all sequences analyzed [5-14].

Through an extensive study the RRM has reached a fundamental conclusion: one RRM characteristic frequency characterizes one particular biological function or interaction. This frequency is related to the biological function provided the following criteria are met:

- One peak only exists for a group of protein sequences sharing the same biological function [14].
- No significant peak exists for biologically unrelated protein sequences.
- Peak frequencies are different for different biological functions.

Once the characteristic frequency for a particular protein function/interaction is identified, it is possible then to utilize the RRM to predict the amino acid regions in the sequence which predominantly contributed to this frequency and consequently to the observed function. For that purpose, we used the short term Fourier transformed approach for a coarse-grained "hot spots analysis".

For coarse-grained hot spot analysis, power spectra were estimated from each window resulting from a sliding pattern along the whole protein sequence. The amplitude corresponding to the frequency of interest was estimated for each window. We hypothesize that those windows with larger values of amplitude are those contributing the most to the given function/interaction. The resolution of the method is limited. On one hand the length of the window allows us to mention only "regions of interest" and not specific positions that are crucial for the function. On the other hand, the short duration of the window makes the frequency precision blurry. However, the fact that the method is straightforward as well as the possibility (in the specific case of MYBPC) to compare theoretical predictions with the details available about the protein, provide motivation to test RRM predictions using this approach.

To coarse-grained hot spots analysis we submitted human MYBPC sequences with diverse species and tissue origins.

In order to find out specific frequencies corresponding to different putative interactions/functions, we divided the sequences into the following groups

Group 1: MYBPC from slow muscle

>Q569K7 | Q569K7_HUMAN Myosin binding protein C, slow type - Homo sapiens (Human)

>Q5U3B1 | Q5U3B1_DANRE Myosin binding protein C, slow type - Danio rerio (Zebrafish)

>Myosin-binding protein C, skeletal muscle slow isoform RAT

>Q00872 | MYPC1_HUMAN Myosin-binding protein C, slow-type - Homo sapiens (Human)

Group 2 MYBPC from fast muscle

>P16419 | MYPC2_CHICK Myosin-binding protein C, fast-type - Gallus gallus (Chicken)

>Q5XKE0 | MYPC2_MOUSE Myosin-binding protein C, fast-type - Mus musculus (Mouse)

>Q14324 | MYPC2_HUMAN Myosin-binding protein C, fast-type - Homo sapiens (Human)

>A1L4G9 | A1L4G9_HUMAN Myosin binding protein C, fast type - Homo sapiens (Human)

Group 3 MYBPC from cardiac muscle

>Q14896 | MYPC3_HUMAN Myosin-binding protein C, cardiac-type - Homo sapiens (Human)

>Q676A1 | Q676A1_OIKDI Cardiac myosin-binding protein C - Oikopleura dioica (Tunicate)

>Q90X86 | Q90X86_XENLA Cardiac myosin-binding protein C - Xenopus laevis (African clawed frog)

>Q2Q1P6 | Q2Q1P6_CANFA Cardiac myosin binding protein C - Canis familiaris (Dog)

>Q0VD56 | Q0VD56_BOVIN Myosin binding protein C, cardiac - Bos taurus (Bovine)

>O70468 | MYPC3_MOUSE Myosin-binding protein C, cardiac-type - Mus musculus (Mouse)

>A5PL00 | A5PL00_HUMAN Myosin binding protein C, cardiac - Homo sapiens (Human)

>A2AGQ1 | A2AGQ1_MOUSE Myosin binding protein C, cardiac - Mus musculus (Mouse)

>Q90688 | MYPC3_CHICK Myosin-binding protein C, cardiac-type - Gallus gallus (Chicken)

Results and discussion.

In figure 1 the results of RRM analysis for groups 1 and 2 separately are represented. In both groups a large, prominent peak appear at f=0.3667, but also other smaller peaks are present. The presence of several peaks might reflect the presence of more than one function/interaction among the proteins of this group, or might be the result of nonsense fluctuations in obtained spectra given the limited set of sequences included into the analysis.

When only cardiac MYBPC are considered, the peak at 0.3667 remains, but it is not the largest one. The peak at frequency 0.226 is the most prominent (figure 2).

Apparently this peak at f=0.22 corresponds to the cardiac variant of MyBPC.

Table 1 summarizes the results obtained for each peak

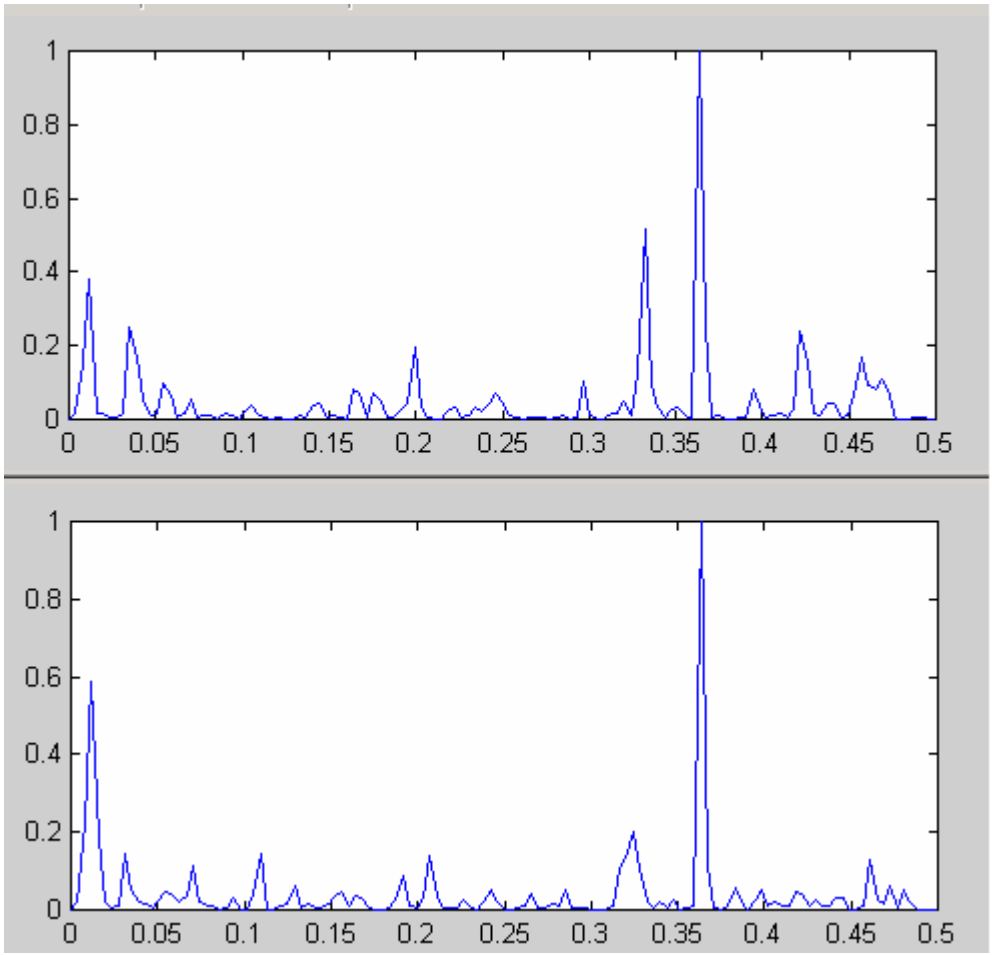

Figure 1. Cross spectra obtained for slow (above) and fast muscle isoforms of MyBPC. For details about the sequences included see methods.

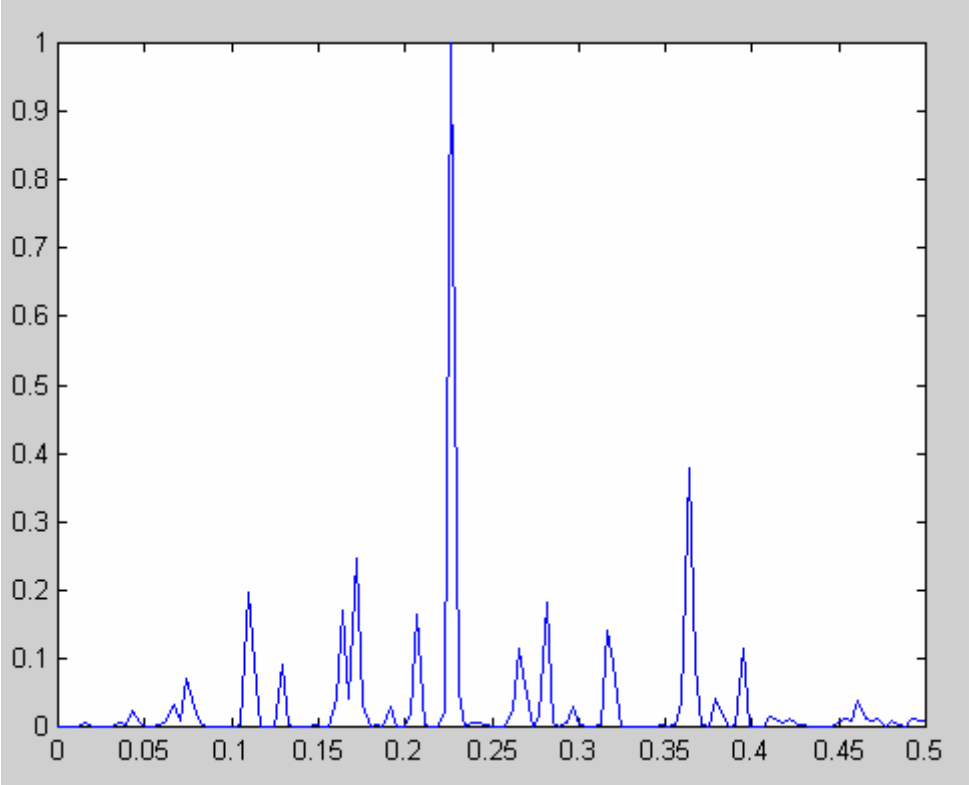

Figure 2. Cross spectrum obtained for cardiac isoforms of MYBPC. For details, see figure 1.

In the following section, we explore the results of short Fourier transformed analysis on cardiac MYBPC for the frequency 0.226.

As it can be noticed from figure 3,

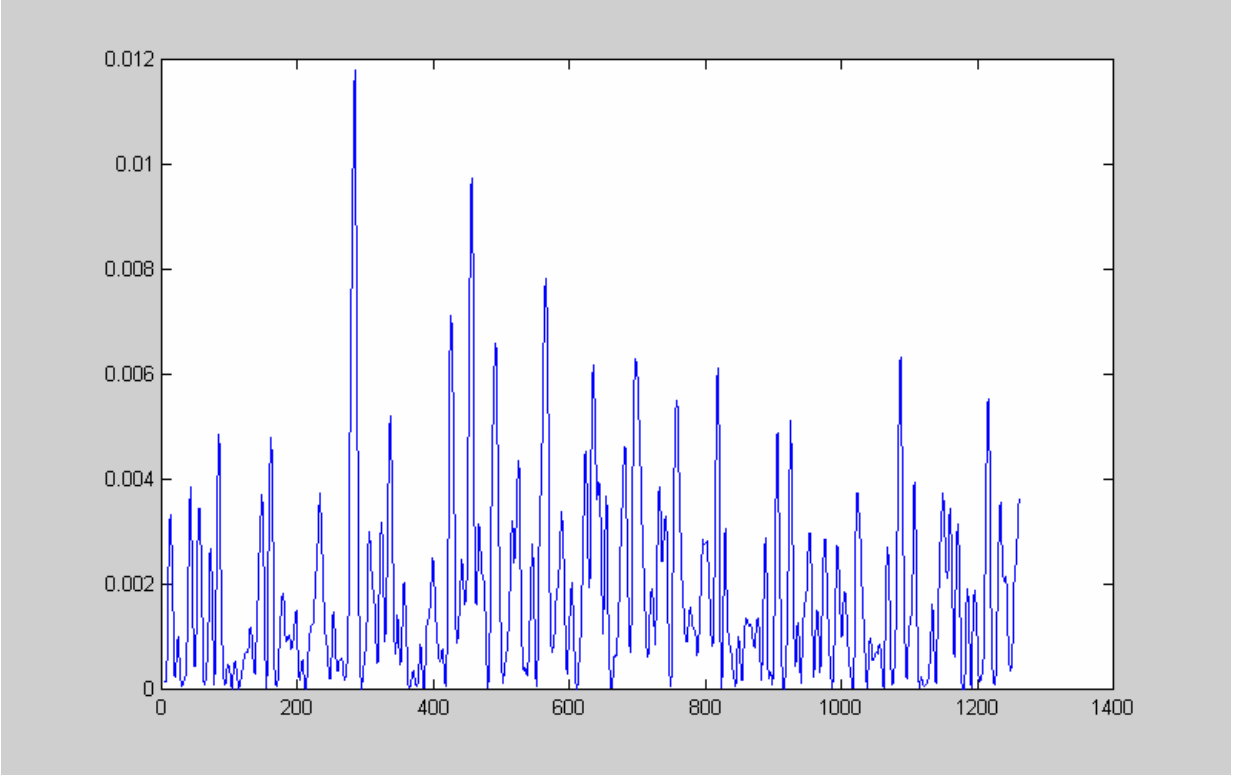

Figure 3. "Contribution" of MYBPC segments to the spectral peak at f=0.22. Abscissas: Position in the protein sequence. Ordinates: Effect of suppression of f=0.22 into the spectral density. Each point corresponds to a 19-AA window centered at the position corresponding to the abscissa value.

The most sensitive region for the peak at 0.226 is a window of 19 amino acids centered at position 284.

Other positions are also outstanding in figure 3, especially in the region spanning from AA 400 and 800, which roughly corresponds to domains C3-C5 of the protein. In the following section, an attempt will be made to correlate the findings from figure 3 with available knowledge about cardiac MYBPC.

A similar analysis was carried out for frequency at f=0.3667. The most relevant positions were found at windows around 603, 720, and 766

Discussion

On this section, we are attempting to contrast the obtained results with available information about different forms of MYBPC.

Peak at f=0.226

According to literature data, position at 284 is very important for cardiac MYBPC. This site is within the C1-C2 link, which includes a cardiac specific insert. In this region, three phosphorylation sites have been described (unlike the skeletal type, which presents

only one phosphorylation site). As stated the phosphorylation site (S284) is located inside a cardiac specific sequence (LAGGGRRIS sequence, spanning from 276 until 284). In that area is located the first phosphorylation site, where the first phosphate must be added by a CaM-II kinase to residue S284, to make the other phosphorylation sites accessible. Neighboring this site, the following mutations associated to CMH4 have been reported [15-16]:

```
VARIANT    278     278   1    G -> E (in CMH4). VAR_019891
VARIANT    279     279   1    G -> A (in CMH4). VAR_019892
VARIANT    282     282   1    R -> W (in CMH4). VAR_029397
```

Thus this zone, as predicted by the RRM analysis; points to a very important region in the cardiac isoform, indirectly supporting the idea of f=0.22 as corresponding to the specific function of the cardiac isoform.

On the other hand, close to 456 (beginning of C3, which spans from AA452 until AA 542) two mutations associated to CMH4 have been reported:

```
VARIANT    450     450   1    E -> Q (in CMH4). VAR_027879
VARIANT    457     457   1    R -> H (in CMH4). VAR_029399
```

Additionally, a mutation with high lethal rate (4 deaths out of 20 patients with the mutation) has been reported at glu451gln [16]. In table 2, the phosphorylation sites and/or disease causing mutations in the vicinity of relevant zones for frequency f=0.22, are listed. As cutting off criterion, those peaks in figure 3 with amplitudes equal or above 50% of the maximum amplitude were considered.

Position	Amplitude (relative units)	%	Type of mutation/interaction reported
275	11,8	100	Phosphorylation
284	11,8	100	Phosphorylation
278	11,8	100	g-e
279	11,8	100	g-a
282	11,8	100	r-w
450	9,7	82,20339	e-q
457	9,7	82,20339	r-h
565	7,8	66,101695	c-r
489	6,6	55,932203	g-r
494	6,6	55,932203	r-q
495	6,6	55,932203	r-g
809	6,1	51,694915	r-h
810	6,1	51,694915	k-r
819	6,1	51,694915	r-q
754	5,9	50	n-k (destabilizes the structure of domain C5).
769	5,9	50	d-n

Table 2. Comparison of some reported mutations with areas predicted by coarse-grained hot spot analysis. Mutations dwelling inside the corresponding predicted windows have been

included. As cut-off criterion, the corresponding window contributes with a peak higher than 50% of the maximal amplitude.

It is worth of notice that the cardiac specific insert with the most important phosphorylation site and the particularly destructive mutation reported position 754 are associated to this frequency

Peak at f=0.3667

Analysis for f=0.3667 suggests important regions close to the AA position 603. This first region corresponds to the final end of C4, where putatively the angle of the protein changes and the ring around the myosin filament is formed [3]. In the neighborhood of AA603 the following disease-related mutations have been reported:

VARIANT	603	603	1	D -> V (in CMH4).	VAR_029405
VARIANT	604	604	1	D -> N (in CMH4; pathogenicity remains to be determined).	VAR_029406
VARIANT	607	607	1	P -> L (in CMH4).	VAR_029407

The other regions, at 720 and 766, both in c5, a particularly important domain, forming the MYBPC ring around Myosin. These two positions are flanking the cardiac specific insert in C5.

The following mutation has been reported close to the site 766

VARIANT	769	769	1	D -> N (in CMH4).	VAR_029411

Further evidences supporting f=0.22 as associated to cardiac MYBPC

Literature reports 80 nonsense mutations leading to HCM4. We generated an artificial sequence in which EllP values are identical to those from native human cardiac MYBPC, except for the points where each of the 80 mutations were reported. We estimated the cross-spectrum for a collection of cardiac sequences, where the human MYBPC was substituted for the "pathological" protein as designed in our description.

We obtained that the peak at 0.22 was reduced to 69% of its initial value. This suggest that at least an important part of CMH4 associated mutations in MYBPC lead to a reduction in the amplitude at f=0.22

Figure 5 schematically summarizes our results and their contrasting to literature data.

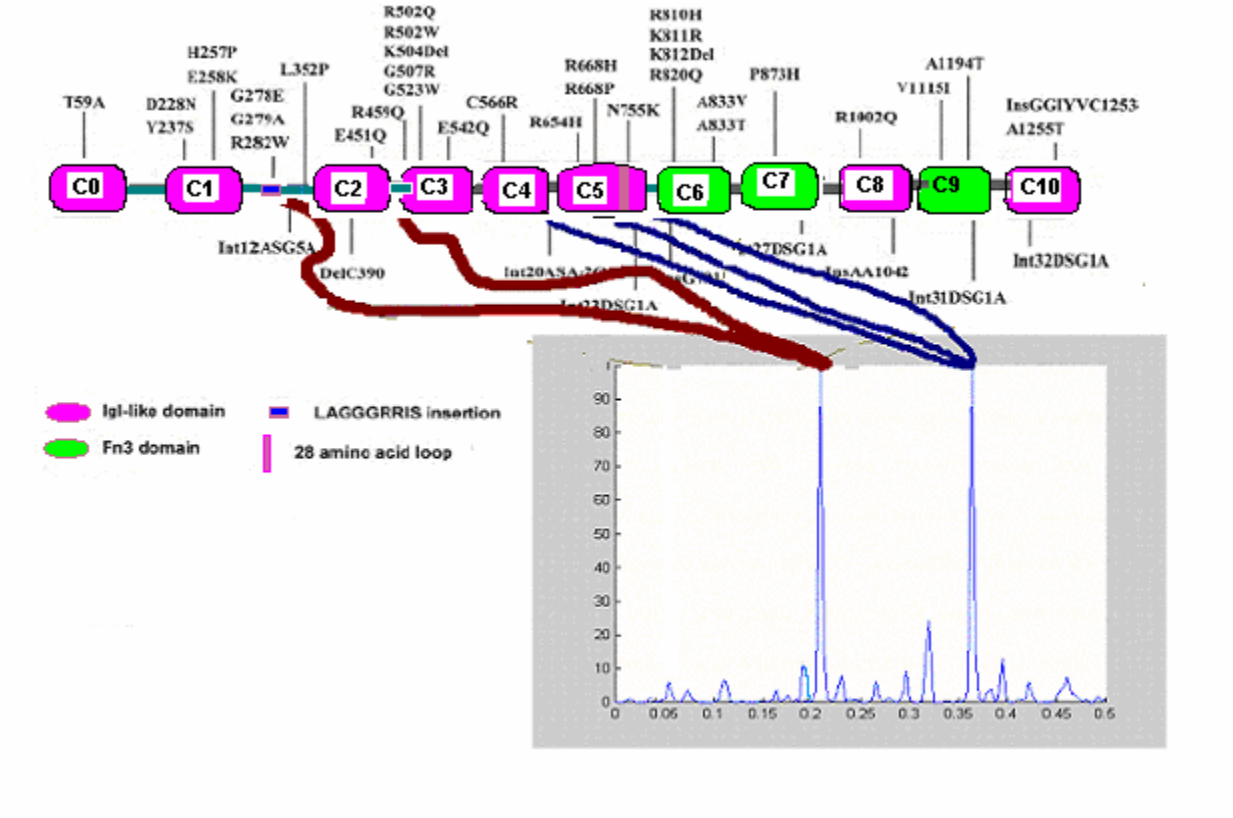

Figure 4. Illustration of the correspondence between frequencies predicted for MYBPC, the hot spots areas predicted by the coarse grained analysis and the most important disease-causing mutations or relevant sites so far described for MYBPC.

Our results indicate that application of RRM analysis to MYBPC yields meaning bearing results when we compare them on the light of known data about this protein and disease-causing mutations in it.

Thus we conclude that at least two RRM frequencies are associated to MYBPC function, being f=0.226 the most specific for cardiac function. Nearly 15% of the point mutations purportedly associated to CMH4, a cardiac specific insert; two phosphorylation sites and a mutation leading to an irreversible denaturation of the dominion C5 fall into the regions of major contribution to f=0.22. An artificial "sequence" with all point mutations reported in literature leads to a reduction of the peak at f=0.22.

We hope that by means of a method similar to the one used in this paper might pave the way for prediction the degree of pathogenicity of real or putative mutations in different proteins. Strategies for disease management based on RRM could be devised.

References

[1] Oakley CE , Hambly BD, Curmi P M and Brown LJ. Myosin binding protein C: Structural abnormalities in familial hypertrophic cardiomyopathy. Cell Research (2004) 14, 95–110

[2] Maron BJ. Triggers for sudden cardiac death in the athlete. Cardiol Clin 1996; 14(2):195–210.

[3] Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT, Bild DE. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. Circulation 1995; 92(4):785

[4] Miura K, Nakagawa H, Morikawa Y, *et al*. Epidemiology of idiopathic cardiomyopathy in Japan: results from a nationwide survey. Heart 2002; 87(2):126–30

[5] Cosic I., (1997) The Resonant Recognition Model of Macromolecular Bioactivity: Theory and Applications, Birkhauser Verlag, Basel, Switzerland.

[6] Ciblis P., Cosic I., (1997) The Possibility of Soliton/Exciton Transfer in Proteins, Journal of Theoretical Biology, 184, 331-338.

[7] De Trad C. H., Fang Q., Cosic I., (2000), "The Resonant Recognition Model (RRM) Predicts Amino Acid Residues in Highly Conservative Regions of the Hormone Prolactin (PRL)", Biophysical Chemistry, 84/2, 149-157.

[8] Cosic I., (2001), "The Resonant Recognition Model of Bio-molecular Interactions: possibility of electromagnetic resonance", Polish Journal of Medical Physics and Engineering, vol. 7. No.1, 73-87

[9] Pirogova, E., Fang, Q., Akay, M., Cosic, I., (2002) "Investigation of the structure and function relationships of Oncogene proteins", Proceeding of the IEEE, Vol. 90, Issue 12, pp. 1859-1867.

[10] De Trad C. H., Fang Q., Cosic I., (2002), "Protein sequences comparison based on the wavelet transform approach", Protein Engineering, Vol.15, no.3, pp 193-203.

[11] Pirogova, E., Simon, G.P., Cosic, I., (2003) "Investigation of the applicability of Dielectric Relaxation properties of amino acid solutions within the Resonant Recognition Model", IEEE Transactions on NanoBioscience, Vol.2, Issue 2, pp.63-69.

[12] Cosic I., (1994) Macromolecular Bioactivity: Is it Resonant Interaction between Macromolecules? - Theory and Applications, IEEE Trans. on Biomedical Engineering, 41, 1101-1114.

[13] Cosic I., Drummond A.E., Underwood J.R., Hearn M.T.W., (1994) A New Approach to Growth Factor Analogue Design: Modelling of FGF Analogues, Molecular and Cellular Biochemistry, 130, 1-9.

[14] Cosic I., (1995) Virtual Spectroscopy for Fun and Profit, Biotechnology, 13, 236-238.

[15] Moolman-Smook J, Flashman E, de Lange W , et al. (2002) Identification of novel interactions between domains of Myosin binding protein-C that are modulated by hypertrophic cardiomyopathy missense mutations. Circ Res 2002; 91(8):704–11.

[16] Niimura H, Bachinski LL, Sangwatanaroj S, Watkins H, Chudley AE, McKenna W, et al.. Mutation in the gene for cardiac myosin-binding protein C and late-onset familial hypertrophic cardiomyopathy. N Engl J Med. 1998;338:1248-57.