ELSEVIER

Biophysical Chemistry 84 (2000) 149–157

Biophysical Chemistry

www.elsevier.nl/locate/bpc

The resonant recognition model (RRM) predicts amino acid residues in highly conserved regions of the hormone prolactin (PRL)

Chafia Hejase de Trad, Qiang Fang, Irena Cosic*

Department of Electrical and Computer Systems Engineering, Monash University, Wellington Road, Clayton, VIC 3168, Australia

Received 8 October 1999; accepted 22 December 1999

Abstract

The resonant recognition model (RRM) is a model which treats the protein sequence as a discrete signal. It has been shown previously that certain periodicities (frequencies) in this signal characterise protein biological function. The RRM was employed to determine the characteristic frequencies of the hormone prolactin (PRL), and to identify amino acids ('hot spots') mostly contributing to these frequencies and thus proposed to mostly contribute to the biological function. The predicted 'hot spot' amino acids, Phe-19, Ser-26, Ser-33, Phe-37, Phe-40, Gly-47, Gly-49, Phe-50, Ser-61, Gly-129, Arg-176, Arg-177, Cys-191 and Arg-192 are found in the highly conserved amino-terminal and C-terminus regions of PRL. Our predictions agree with previous experimentally tested residues by site-direct mutagenesis and photoaffinity labelling.

Keywords: Prolactin; Resonance recognition model; Continuous wavelet transform; Binding site

1. Introduction

Prolactin (PRL) is a polypeptide hormone of approximately 200 amino acid residues (MW ~ 23 000), with six cysteines, three interchain disulfide bridges and a half-life ($t_{1/2}$) of approximately 10 min [1]. It is synthesised as a 227-amino-acid precursor protein which has 29 residues cleaved off to generate the active hormone. It is structurally homologous to growth hormone and to placental hormone chorionic somatomammotropin (hCS), which has both milk-producing and milk-promoting activity [2,3]. These three

* Corresponding author Tel.: +61-3-9905-5356; fax: +61-3-9905-3454.
E-mail address: irena.cosic@eng.monash.edu.au (I. Cosic)

0301-4622/00/$ - see front matter
PII: S0301-4622(00)00109-5

hormones have evolved from a single ancestral gene.

PRL is involved in more than 85 biological functions in vertebrates. It is a promoter of mammary gland growth and milk production in a variety of mammals [4]. It may also be involved in the growth, differentiation and maturation of cells of the immune system [5]. Because of the wide spectrum of functions, the hormonal signal might be mediated by specific membrane receptors [6] and might be regulated by nucleotides [7–9].

Several studies aimed at determining which region of PRL interacts with the receptor have been reported. Studies like sequence comparison [10] and chemical modifications [10–14] have characterised some important structural residues with biological importance to PRL. It has been suggested that PRL binds to two molecules of prolactin receptor (PRLR) through two regions known as binding sites 1 and 2 [15]. Also a nucleotide binding site has been characterised on ovine PRL and the peptide Ala-22 to Tyr-28 identified near the N-terminus was suggested as being involved in this binding [9].

An alternative way to analyse protein–protein and protein–DNA interactions has been introduced [16,17]. This approach, known as the resonant recognition model (RRM), interprets the protein linear information using signal analysis methods [16–25]. It has been shown that certain periodicities (frequencies) within the distribution of energies of delocalised electrons along the protein are critical for protein biological function (i.e. interaction with its target). Once the RRM characteristic frequency for a particular biological function or interaction has been determined, it is possible to identify the individual amino acids so-called 'hot spots', or domains that contribute mostly to the characteristic frequency and thus to the protein's biological function as well [21–24].

In this paper, we have applied the RRM to the hormone PRL with the aim to find the resonant frequencies, predict functionally important amino acids, 'hot spots', within the protein sequence and then compare them to the experimentally proposed amino acids by most researchers.

2. Methods

The RRM is a physical and mathematical model which interprets protein sequence linear information using signal analysis methods [16–20]. It comprises two stages. The first involves the transformation of the amino acid sequence into a numerical sequence. Each amino acid is represented by the value of the electron–ion interaction potential (EIIP) [26] which describes the average energy states of all valence electrons in a particular amino acid. The EIIP values for each amino acid were calculated using the following general model pseudopotential [26]:

$$\langle k+q|w|\,k\rangle = 0.25Z\sin(\pi 1.04Z)/(2\pi) \qquad (1)$$

where q is a change of momentum of the delocalised electron in the interaction with potential w, while:

$$Z = (\Sigma Z_i)/N \qquad (2)$$

where Z_i is the number of valence electrons of the ith component of each amino acid and N is the total number of atoms in the amino acid. Each amino acid or nucleotide, irrespective of its position in a sequence, can thus be represented by a unique number.

Numerical series obtained this way are then analysed by digital signal analysis methods in order to extract information pertinent to the biological function. The original numerical sequence is transformed to the frequency domain using the discrete Fourier transform (DFT). As the average distance between amino acid residues in a polypeptide chain is approximately 3.8 Å, it can be assumed that the points in the numerical sequence derived are equidistant. For further numerical analysis the distance between points in these numerical sequences is set at an arbitrary value $d = 1$. Then the maximum frequency in the spectrum is $F = 1/2d = 0.5$. The total number of points in the sequence influences the resolution of the spectrum only. Thus for N-point sequence the resolution in the spectrum is equal to $1/N$. The nth point in the spectral function corresponds to the frequency $f = n/N$.

In order to extract common spectral characteristics of sequences having the same or similar biological function, the following cross-spectral function was used:

$$S_n = X_n Y_n^* \qquad n = 1,2,\ldots,N/2 \tag{3}$$

where X_n are the DFT coefficients of the series $x(m)$ and Y_n^* are complex conjugate DFT coefficients of the series $y(m)$. Peak frequencies in the amplitude cross-spectral function define common frequency components of the two sequences analysed.

To determine the common frequency components for a group of protein sequences, we have calculated the absolute values of multiple cross-spectral function coefficients M, which are defined as follows:

$$|M_n| = |X1_n| \cdot |X2_n| \ldots |XM_n| \qquad n = 1,2,\ldots,N/2 \tag{4}$$

Peak frequencies in such a multiple cross-spectral function denote common frequency components for all sequences analysed. Signal-to-noise ratio (S/N) for each peak is defined as a measure of similarity between sequences analysed. S/N is calculated as the ratio between signal intensity at the particular peak frequency and the mean value over the whole spectrum. The extensive experience gained from previous research [16–24] suggests that a S/N ratio of at least 20 can be considered as significant. The multiple cross-spectral function for a large group of sequences with the same biological function has been named 'consensus spectrum'. The presence of a peak frequency with significant signal-to-noise ratio in a consensus spectrum implies that all of the analysed sequences within the group have one frequency component in common. This frequency is related to the biological function provided the following criteria are met:

1. one peak only exists for a group of protein sequences sharing the same biological function;
2. no significant peak exists for biologically unrelated protein sequences; and
3. peak frequencies are different for different biological functions.

In our previous studies, the above criteria have been tested with over 1000 proteins from 25 functional groups [17–25]. The following fundamental conclusion was drawn from our studies: each specific biological function of protein or regulatory DNA sequence(s) is characterised by a single frequency.

Once the RRM characteristic frequency for a particular biological function has been determined, it is possible to identify the individual amino-acids so called 'hot spots' (using FT) or domains (using CWT) that contribute mostly to the characteristic frequency and thus to the protein's biological function as well [16–25].

PRL sequences from different species are transformed into a numerical series using the electron–ion interaction potential (EIIP) for each amino acid [26]. These new series are then transformed into the frequency domain using Fourier transform methods. A multiple cross-spectral function of all these series is generated to determine the common frequency components for that group of PRL sequences [16–19]. The multiple cross-spectral function is analysed for characteristic peaks with significant signal-to-noise (S/N) ratios (S/N ~ 20 is significant). Such peaks are strongly correlated with specific functional groups of the hormone.

Each of the characteristic frequencies is used to determine the positions of amino acids significant to the biological functions of the hormone PRL. The positions of the amino acids that are mostly affected by the change of amplitude at a particular frequency are considered as 'hot spots' for the corresponding biological function [21–24]. This prediction method has been used with other proteins including lysozyme, haemoglobin and interleukin-2 [22,23]. The predicted results for these proteins have documented evidence to corroborate their active and important role in the proteins' functions.

The continuous wavelet transform (CWT) method has been also employed to predict the hormone's critical amino acid sites [25]. With the understanding that the protein's active site is

usually made up of domain(s) within the protein sequence rather than single amino acids, a spatial-frequency distribution of the PRL signal was constructed with the CWT method. This distribution is referred to as a scalogram. Only the highest energy domains are considered and analysed at a particular characteristic frequency.

The wavelet transform (WT) is a relatively new signal processing tool efficient for multi-resolution analysis and local feature extraction of non-stationary signals [27,28]. The wavelet transform can be viewed as an inner product operation that measures the similarity or cross-correlation between the signal and the wavelets. The continuous version of the wavelet transform (CWT) of signal $s(t)$ is defined as:

$$\mathrm{CWT}(a,b) = \int s(t)\frac{1}{(a)^{1/2}}\Psi\frac{(t-b)}{a}\mathrm{d}t \qquad (5)$$

where b is the shift factor (the translation factor of the wavelet function along the time axis) and a is the scale factor (it scales a function by compressing or stretching it).

CWT is one of the time or space–frequency representations. A time (space) frequency representation of a signal provides information about how the spectral content of the signal evolves with time (space), thus providing an ideal tool to dissect, analyse and interpret signals with transients or localised events. This is performed by mapping a one-dimensional signal in the time (space) domain into a two-dimensional time (space)–frequency representation of the signal. Because CWT provides same time/space resolution for each scale and thus, CWT can be chosen to localise individual events, such as the active site identification. The particular wavelet chosen here for critical amino acids identification is the Morlet, which is a locally periodic wavetrain:

$$\omega(t) = C\exp\left(\frac{-t^2}{2} + j\omega_0 t\right) \qquad (6)$$

where $\omega_0 = 5.33$ and C is the constant used for normalisation.

From Eq. (6), it can be seen that the Morlet wavelet is a complex sine wave modulated by a Gaussian function. The time–frequency version of CWT can be achieved by making the substitution $a = f_0/f$

$$cwt(t,f) = \int s(\tau)\left(\frac{f}{f_0}\right)^{1/2}\Psi\left(\frac{f}{f_0}(\tau - t)\right)\mathrm{d}t \qquad (7)$$

in which the analysing wavelet becomes essentially a prototype bandpass filter with centre time $t = 0$ and centre frequency f_0. The centre frequency and frequency bandwidth of the CWT vary with scale. However, their ratio remains fixed. It is the constant property of the wavelet.

The underlying property of wavelets is that they are pretty well localised in both time and frequency [29]. The product of the uncertainties of both time and the frequency is bound by the Heisenberg's uncertainty principle; no filter can have a width product smaller than $1/\pi$. The Gaussian filters (Morlet wavelet) attain this theoretical limit.

Strictly speaking, the CWT provides a time–scale representation rather than a time–frequency representation. However, the scale factor of CWT is closely related to the frequency and this makes the mapping from time–scale representation to time–frequency representation possible. That a short duration event (small scale) is inherently dominated by the high frequency components means that the centre frequency f is inversely proportional to the scale a [30]. This relationship can be expressed by $f = K/a$, where K is the proportionality constant which is determined by the particular wavelet function used. For the Morlet wavelet, $K = 0.8125$, and the scale–frequency mapping relationship is:

$$f = \frac{0.8125}{a} \qquad (8)$$

By using Eq. (8), the centre frequency in each scale a can be determined. However, this scale to frequency conversion is just for the sake of convenience that could bring a straightforward interpretation. The fact is that a scale contains information from one frequency band rather than from one single frequency. The CWT is essentially a time (space)–scale representation.

The active sites and the 'hot spot' amino acids are the local energy maxima in the space–frequency representation of a protein signal. According to the RRM, the possible delocalised charges pass different energy stages along the protein backbone and radiate the electromagnetic wave crucial for biological recognition and interaction [17]. The amino acids with local maxima values of CWT coefficients are the energy most concentrated locations of an energy stage where the changes of the kinematic parameters (velocity, acceleration, etc.) of the delocalised charges are most likely to happen. Because the charge transfer speed is most likely to be changed at the amino acid positions with the wavelet transform modulus maxima, the radiation properties are also essentially affected by these amino acids sites. Thus, those sites are naturally proposed as the protein's biological critical sites.

3. Results and discussion

We applied the RRM method to PRL and its precursor. Cross-spectral analysis of 10 PRL sequences from different origins revealed two common frequency peaks (Fig. 1). Table 1 shows the characteristic frequencies obtained. The frequency $f = 0.0078 \pm 0.005$ is common for both the PRL hormone and its precursor. In a previous report, the frequency $f = 0.293 \pm 0.016$ was assigned with the functional groups of growth factors including growth hormone GH [17]. Our predicted frequency $f = 0.2852 \pm 0.005$ for PRL is compatible with the growth factors (including GH) frequency $f = 0.293$. Since PRL and growth hormone are highly structurally homologous [2,3] the predicted frequency component $f = 0.2852$ can be associated with the growth promoting function of prolactins.

Furthermore, we used both frequencies to determine amino acid residues which might be correlated with the functional groups of PRL. The following amino acid residues, termed as 'hot spots' (Fig. 2), were identified:

- using $f = 0.2852$ (Phe-19, Ser-26, Ser-33, Phe-37, Phe-40, Gly-49, Ser-61, Gly-129); and
- using $f = 0.0078$ (Phe-50, Arg-176, Arg-177, Cys-191, Arg-192).

Table 1
Characteristic frequencies of the hormone prolactin and its precursor obtained from the RRM method

Protein sequence	Frequency (relative units)	Number of sequences	S/N
Prolactin	0.0078	10	161
	0.2852		30
Prolactin precursor	0.0078	10	207

The CWT amino acid identification was also applied to PRL sequence. The wavelet representation (Fig. 2) or scalogram shows a space–frequency distribution of the PRL signal. Only the high-energy domains for a frequency band are considered (bright regions in the scalogram). Note the brightest regions between scales 2 and 3.5, 4.5 and 6.5 and finally 8.5 and 10. The brightest domains are correlated with amino acid residues 10–28, 30–42, 45–55, 60–70, 85–95, 110–130, 170–180 and 182–198. The important residues found by the CWT method could dominate different frequency bands.

Since both human growth hormone and PRL are highly homologous, Goffin et al. [31] developed a three-dimensional model for PRL considered a unique atomic structure [31]. Crystallographic and mutational reports of human growth hormone have led to the identification of the amino acids on binding sites 1 and 2 [32,33]. Binding site 1 of human growth hormone is made up of residues of helices 1 and 4 and loop 1. Binding site 2 is made of residues located on the opposite sides of helices 1 and 3 in addition to a few residues in the small N-terminal loop. To date, no crystallographic structure has been reported for PRL other than Goffin's model [31]. Based on this model, we compared the predicted 'hot spots' to amino acid residues corresponding to binding site 1 (Table 2).

The CWT method predicted the 10–28 region near the N-terminus of PRL to be of biological importance. The peptide region near the N-terminus has been implicated in the biological

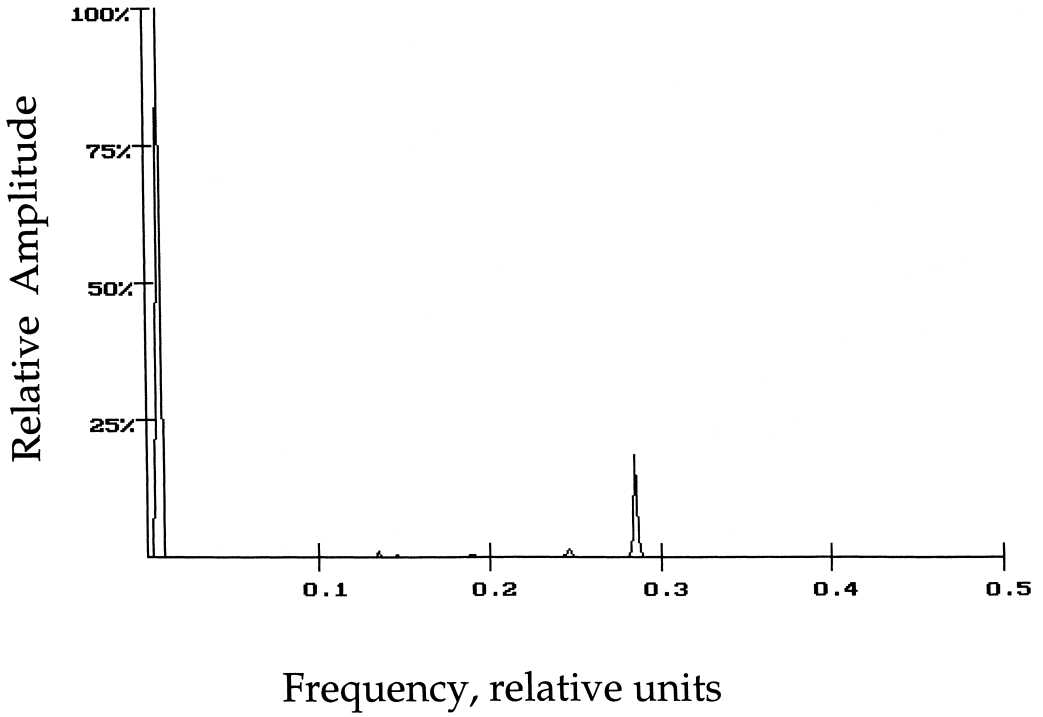

Fig. 1. Multiple cross-spectral function of 10 prolactin sequences. The prominent peaks are located at the frequencies 0.0078 and 0.2852.

activity of PRL [34]. 'Hot spot' analysis revealed the significance of residues Phe-19 and Ser-26. Note that Ser-26 corresponds to the peptide region Ala-22 toTyr-28 previously identified as an NADH/NADPH nucleotide-binding site [9]. Also Ser-26 resides in the middle of the putative first helix of the developed three-dimensional model of PRL [31]. Identification of this region suggests a possible physiological role for this interaction, possibly to induce a conformational change upon binding the first helix. This may lead to dissociation of ligand from its high-affinity receptors upon internalisation, allowing receptors to be recycled [9]. Also, this binding may reflect catalytic or regulatory properties of PRL after internalisation such as exists with NAD^+ binding toxins. Other studies with rhIL-2 (recombinant human interleukin-2), glucagon and rmGM-CSF (recombinant murine granulocyte/macrophage colony stimulating factor) have also identified the peptide regions on these molecules involved in the nucleotide binding and biological activity [35,36]. Thus our theoretical results agree with the experimental findings [9].

The RRM method predicted the amino acid residues Ser-33, Phe-37 and Phe-40 to be of particular importance. One of the high-energy domains on the CWT scalogram involved the 30–42 region in which the predicted residues are subsets. Kinet et al. [37] have shown that the amino acid residues His-30 and Phe-37 are binding determinants of human PRL. However, the other two predicted amino acids Ser-33 and Phe-40 could prove equally important if tested appropriately.

The region 60–70 corresponded to a high-energy domain on the CWT scalogram. In agreement, mutational analysis of the 58–74 peptide segment on loop 1 clearly demonstrated the involvement of this region in PRL bioactivity [15]. We propose that the amino acid Ser-61 be adequately tested since hot spot analysis revealed its significance.

A second receptor binding site on human PRL

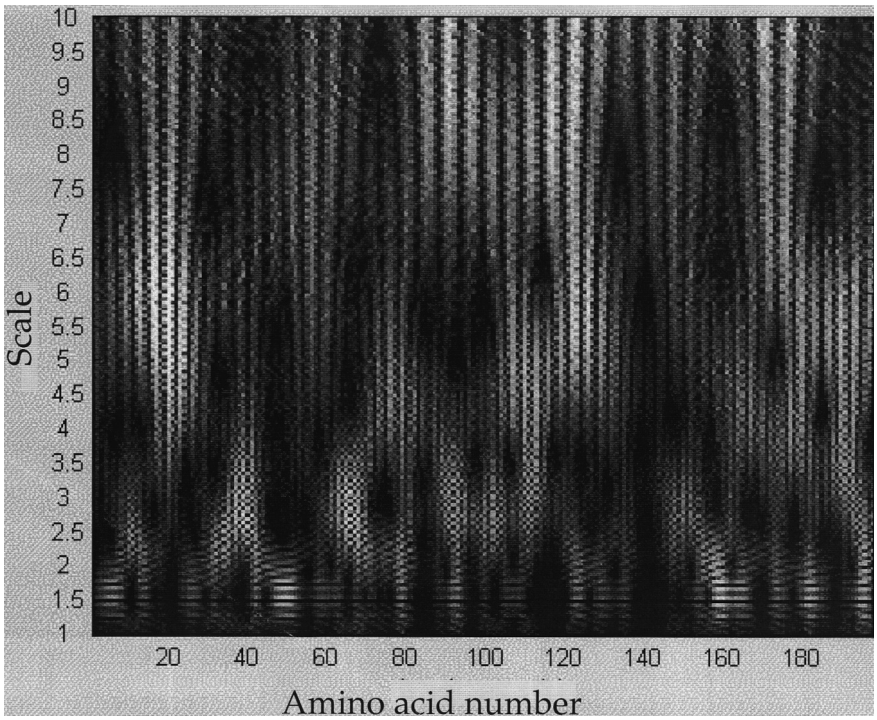

Fig. 2. Continuous CWT scalogram of prolactin. The abscissa represents the position of amino acid residues and the ordinate represents the continuous scales.

has been proposed by Goffin et al. using site-direct mutagenesis [38].

Most of the residues suggested to be potentially involved in the interaction with a second PRLR are in the region delimited by the interface between helices 1 and 3. The steric importance of the residue Gly-129 in maintaining the geometry of binding site 2 was elucidated [38]. Our RRM results also predicted the importance of Gly-129 as a significant amino acid. Other potentially involved residues 110–130 on helix 3 of binding site 2 were also predicted on the CWT scalogram.

The fourth helix (helix 4) in the human PRL hormone plays a role in binding to the PRL receptor through topologically different amino acids in binding site 1 [15,31]. Our results revealed the importance of the regions 170–180 and 182–192. The RRM 'hot spot' analysis identified the amino acid residues Arg-176, Arg-177, Cys-191 and Arg-192 as possible binding determinant candidates. Kinet et al. reported that Arg-176 is a binding determinant on helix 4 [37]. Luck et al. reported that the amino acid residue Arg-177 is unambiguously important for the mitogenic activ-

Table 2
Representation of the amino acid residues corresponding to binding site 1 of the human prolactin model from Goffin et al. [15,31]

Helix 1	Phe-19	Val-23	Ser-26	His-30	Ser-34	Phe-37
Helix 4	Tyr-169	His-173	Arg-176	Arg-177	His-180	Lys-181
	Asn-184	Tyr-185	Leu-188	Arg-192		
Loop 1	His-59	Ser-61	Leu-63	Phe-66	Glu-67	Asp-68
	Lys-69	Glu-70	Glu-71	Glu-73	Lys-75	Asn-76

ity of bovine PRL toward Nb2 cells [12]. Hence, the involvement of this arginine residue in all prolactins proposes this residue a major binding determinant of human PRL.

Keeping in mind that the majority of the theoretically predicted amino acids in this paper are functionally important, we can conclude that this study thus confirms our earlier hypothesis that the RRM method is a novel approach to predict the amino acids that contribute significantly to a protein's biological function. Moreover, our theoretical predictions have shown a good agreement with experimental results. The combined method based on Fourier and wavelet transforms is described here for use in prediction of PRL active sites.

According to the resonant recognition model, the highly selective interaction between protein and its target is based on energy transfer through electromagnetic oscillations of a specific frequency [16,17]. These oscillations could be produced by delocalised charge moving along the proteins' backbone and will pass different energy stages formed by different side chains.

The Fourier approach for active sites identification is through finding the specific residues that mostly affect the RRM characteristic frequency, while the wavelets approach identify the domains of extrema moduli of wavelet coefficients. The domains with local extreme value of WT coefficient indicate the sharpest variation locations of energy stages and those locations have the highest probability to affect the delocalised charge's moving and radiating conditions. Consequently, those domains are naturally proposed as the protein's functional and structural active sites.

References

[1] M. Wallis, in: B. Weinstein (Ed.), Chemistry and Biochemistry of Amino Acids, Peptides and Proteins, vol. 5, Decker, New York, 1978, pp. 213–320.

[2] I.A. Forsythe, Oxford Rev. Reprod. Biol. 13 (1991) 97–148.

[3] A.G. Frantz, in: J.A. Parsons (Ed.), Peptide Hormones, Univ Park Press, Baltimore, 1976, pp. 199–231.

[4] I.A. Forsythe, in: T.B. Mempham (Ed.), Biochemistry of Lactation, Elsevier, Amsterdam, 1983, pp. 309–349.

[5] L.Y. Yu-Lee, Proc. Soc. Exp. Biol. Med. 215 (1997) 35–52.

[6] C. Bole-Feysot, V. Goffin, M. Edery, N. Binart, Endocr. Rev. 19 (1998) 225–268.

[7] R.L. Potter, B. Haley, Methods Enzymol. 91 (1983) 613–633.

[8] A.J. Chavan, Y. Nemoto, S. Narumiya, S. Kozaki, B.E. Haley, J. Biol. Chem. 267 (1992) 14866–14870.

[9] C.H. Trad, A.J. Chavan, J. Clemens, B.E. Haley, Arch. Biochem. Biophys. 304 (1993) 58–64.

[10] C.S. Nicoll, G.L. Mayer, S.M. Russell, Endocr. Rev. 7 (1996) 169–203.

[11] P.C. Necessary, T.T. Andersen, K.E. Ebner, Mol. Cell. Endocrinol. 39 (1985) 247–254.

[12] D.N. Luck, M. Huyer, P.W. Gout, C.T. Beer, M. Smith, Mol. Endocrinol. 5 (1991) 1880–1886.

[13] D.N. Luck, P.W. Gout, C.T. Beer, M. Smith, Mol. Endocrinol. 3 (1989) 822–831.

[14] D.N. Luck, P.W. Gout, K. Kelsay, T. Atkinson, C.T. Beer, M. Smith, Mol. Endocrinol. 4 (1990) 1011–1016.

[15] V. Goffin, M. Norman, J.A. Martial, Mol. Endocrinol. 6 (1992) 1381–1392.

[16] I. Cosic, The Resonant Recognition Model of Macromolecular Activity, Birkhaouser, 1997.

[17] I. Cosic, IEEE Trans. Biomed. Eng. 41 (1994) 1101–1114.

[18] I. Cosic, in: D. Wise (Ed.), Bioinstrumentation and Biosensors, Marcel Dekker, New York, 1990, pp. 475–510.

[19] I. Cosic, Bio/Technology 13 (1995) 236–238.

[20] I. Cosic, Med. Biol. Eng. Comp. 34 (1996) 139–140.

[21] I. Cosic, D. Nesic, Eur. J. Biochem. 170 (1988) 247–252.

[22] I. Cosic, V. Pavlovic, V. Vojisavljevic, Biochimie 71 (1989) 333–342.

[23] I. Cosic, A. Hodder, M. Aguilar, M.T.W. Hearn, Eur. J. Biochem. 198 (1991) 113–119.

[24] I. Cosic, M.T.W. Hearn, J. Mol. Recognition 4 (1991) 57–62.

[25] Q. Fang, I. Cosic, APESEM 21 (1998) 179–185.

[26] V. Veljkovic, I. Slavic, Phys. Rev. Lett. 29 (1972) 105–108.

[27] I. Daubechies, Commun. Pure Appl. Math. 41 (1988) 909–996.

[28] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, 1992.

[29] G. Strang, T. Nguyen, Wavelets and Filter Banks, Cambridge Press, Wellesley, 1996.

[30] M. Vetterli, C. Herley, IEEE Trans. Signal Proc. 40 (1992) 2207–2232.

[31] V. Goffin, J.A. Martial, N.L. Summers, Protein Eng. 8 (1996) 1215–1231.

[32] B.C. Cunningham, J.A. Wells, Science 244 (1989) 1081–1084.

[33] A.M. de Vos, M. Ultsch, A.A. Kossiaboff, Science 255 (1992) 306–312.

[34] C. Clapp, R.I. Weiner, Endocrinology 130 (1992) 1380–1386.
[35] M. Shoemaker, P.C. Lin, B. Haley, Protein Sci. 1 (1992) 884–891.
[36] M.A. Doukas, A.J. Chavan, C. Gass, T. Brone, B. Haley, Bioconjugate Chem. 3 (1992) 484–492.
[37] S. Kinet, V. Goffin, V. Mainfroid, J. Martial, J. Biol. Chem. 271 (1996) 14353–14360.
[38] V. Goffin, I. Struman, V. Mainfroid, S. Kinet, J. Martial, J. Biol. Chem. 269 (1994) 32598–32606.