Chem-Bio Informatics Journal, Vol. 7, No. 2, pp.35-48 (2007)

Nuclear localization of proteins with a charge periodicity of 28 residues

Noriyuki SAKIYAMA*, Runcong KE, Ryuusuke SAWADA
Masashi SONOYAMA and Shigeki MITAKU

Department of Applied Physics, Graduate School of Engineering, Nagoya University
Furocho, Chikusa-ku, Nagoya 464-8603, Japan

**E-mail: sakiyama@bp.nuap.nagoya-u.ac.jp*

(Received: October 10, 2007; accepted: November 26, 2007; published online: December 19, 2007)

Abstract

Proteins with a charge periodicity of 28 residues (PCP28) were found recently in the human proteome, and many of the annotated PCP28 were located in the nucleus (Ke *et al.*, Jpn. J. Appl. Phys. 2007). The physical properties of the amino acid sequences were analyzed to detect the difference in the physicochemistry between the nuclear and cytoplasmic PCP28 and develop a software system to classify the two types of PCP28. A significant difference in the global parameters from the entire sequence and the local parameters around a segment with the highest positive charge density was found between the nuclear and cytoplasmic PCP28. The global classification score included the densities of proline and cysteine, and the negative charge density, while the local score included the symmetry of the charge distribution, the density of cysteine, and the positive charge density. A prediction system was developed using the global and local scores, which possessed a sensitivity and specificity of 92% and 88%, respectively. The mechanism of translocation of proteins to the nucleus is discussed using the parameters relevant to the predictive system.

Key Words: proteome, charge distribution, nuclear localization protein, prediction, bioinformatics

Area of Interest: Bioinformatics and Bio computing

Chem-Bio Informatics Journal, Vol. 7, No. 2, pp.35-48 (2007)

1. Introduction

Recent analyses of the distribution of various physical properties in amino acid sequences have revealed several categories of proteins that have a recognizable distribution of physical properties. For example, membrane proteins contain highly hydrophobic regions in their amino acid sequences, corresponding to the transmembrane helices. Combining this feature with clusters of amphiphilic residues at the ends of the helix, a membrane protein predictor SOSUI was developed with an accuracy greater than 95%[1,2]. This indicates that unknown proteins can be classified by the analysis of amino acid sequences alone in terms of several physical parameters[3-5].

Recently, we found that approximately 3% of all amino acid sequences from the human genome show a significant charge periodicity of 28 residues[6]. The largest fraction of proteins with a charge periodicity of 28 residues (PCP28) was nuclear proteins, although many PCP28 were poorly identified. Another category of PCP28 was motor proteins, which have long extended supercoil structures located in the cytoplasm. In fact, many types of proteins belong to the PCP28 group, but the fraction of nuclear PCP28 was larger than the fraction that included all other types of PCP28.

It is well known that the nuclear localization signal in nuclear proteins is not decisive and the sensitivity of various methods for predicting nuclear proteins is not high. This fact suggests that very short segments as the nuclear localization signals are not enough for the translocation of proteins into the nucleus. One of the possible characteristics of proteins for being translocated into the nucleus is the property of longer segments or entire sequences. The majority of PCP28 are nuclear proteins and hence, investigating the properties of PCP28, and in particular comparing the nuclear and the cytoplasmic PCP28, may be useful for obtaining important information about the nuclear localization of proteins.

Here, we first extracted PCP28 from all amino acid sequences in the public database of Swiss-Prot[7]. Since the data from Swiss-Prot are annotated in principle, we classified PCP28 from the Swiss-Prot into two categories: proteins in the nucleus and those in the cytoplasm. Second, the physicochemical properties of the two PCP28 categories were compared, allowing calculation of two discrimination scores from entire amino acid sequences as well as from the local regions around clusters of positive charges (which are characteristic of nuclear localization signals)[8,9]. Using the scores from the global and local parameters, the prediction system SOSUIpcp28nl was developed with a sensitivity of 92% and specificity of 88%.

2. Methods

2.1 Dataset of PCP28 from Swiss-Prot database

The autocorrelation function of the charge distribution of all amino acid sequences from Swiss-Prot database release 48.7 were calculated using the following equation:

$$C(j)=\frac{\sum_{k=1}^{N}\sum_{i=1}^{L(k)-j}[q(i)q(i+j)]}{\sum_{k=1}^{N}[L(k)-j]} \qquad (1)$$

$$(L(k)>j)$$

where $q(i)$ is the electric charge of the amino acid at the i-th position in the sequence, N is the

total number of proteins, $L(k)$ represents the length of the *k*-th protein, and j is the interval of the amino acid sequence. Positively charged residues (+1) include Lys, Arg, and His; negatively charged residues (-1) include Asp and Glu. A similar analysis was also carried out in which we assumed that histidines were not ionized. We observed that the results did not change much, probably owing to the fact that the ratio of histidine was small. The PCP28 was selected according to the method described by Ke *et al*. Comparison of the autocorrelation function at an interval of 28 residues with the corresponding values in the interval region between 14 and 42 residues defined PCP28 as proteins whose autocorrelation function for the interval of 28 residues was larger than other intervals by more than 0.01.

Table 1 lists the subcellular locations of PCP28 in eukaryotic proteomes from Swiss-Prot (proteins whose sequence homology is higher than 25% are omitted in this list). It is remarkable that greater than half of the PCP28 are localized in the nucleus. The classification of nuclear proteins in general is difficult because there are no good sequence motifs that include most nuclear proteins. However, the nuclear proteins are well extracted into the PCP28. The nuclear PCP28 are the subset of the total nuclear proteins, but it is expected that the difference between the nuclear PCP28 and the cytoplasmic PCP28 is more clearly observable than the difference between the mixtures of various nuclear and cytoplasmic proteins. Here we hypothesized that the nuclear localization of PCP28 is due to one of the multiple mechanisms and that its discrimination is much simpler than the discrimination of total nuclear proteins. For developing a prediction system of nuclear PCP28, we prepared a training dataset including data on 29 proteins in the nucleus and the cytoplasm at random from the data in Table 1. All PCP28 were used for testing the performance of the predictor. About a half of PCP28 were nuclear proteins, a quarter were the proteins in the outside of the nucleus. We also carried out a cross validation test using 30 positive and 30 negative data for training and remaining data for testing.

Table 1. Dataset of PCP28 in eukaryotic cells classified into groups of proteins localized in the nucleus or in other locations.

Subcellular location	Number of PCP28	Total Number
Nuclear proteins	158	178
Nucleus & cytoplasm	10	
Ribosome	10	
Cytoplasm	33	75
Other proteins	42	
Unknown proteins	100	100

PCP28 in eukaryotic cells selected from Swiss-Prot database rel.48.7.

Table 2. List of data used for analysis

(A) is nuclear PCP28, (B) is cytoplasmic PCP28 and (C) is other PCP28.

A.

AEF1_DROME	AGL8_SOLTU	AP32_ASAEU	ATF7_PONPY	BCL6B_MOUSE	CHE1_CAEEL
COL14_ARATH	CPC1_CRYPA	CTCF_RAT	DDX23_HUMAN	DHX15_HUMAN	DIMH_HUMAN
EDF1_BRARE	EGR1_CHICK	EGR1_HUMAN	EGR3_HUMAN	ERCC3_MOUSE	FYV7_DEBHA
GAT21_ARATH	GLAS_DROME	GLI4_HUMAN	GSP2_YEAST	H1X_HUMAN	HKR3_HUMAN
HM12_CAEEL	HM19_XENLA	HME2A_XENLA	HME30_APIME	HMEN_ARTSF	I20L2_HUMAN
KLF1_MOUSE	KLF7_HUMAN	KRUP_DROME	LSM4_MOUSE	MAL33_YEAST	MRT4_HUMAN
MSL1_YEAST	MYBA_CHICK	MYB_CHICK	NEK2_HUMAN	NELFB_HUMAN	NIPA_MOUSE
NKX32_MOUSE	NOP14_SCHPO	NUF1_YEAST	POK_DROME	PRDM5_HUMAN	PRDM9_HUMAN
RAD1_YEAST	REXO4_CANAL	RPP38_MOUSE	RRP4_YEAST	RTS2_YEAST	RU2A_CANGA
SRP1_SCHPO	SWC5_GIBZE	TOA2_YEAST	UTP11_HUMAN	UTP11_YEAST	VDR_MOUSE
WOX13_ARATH	WRK26_ARATH	XPA_XENLA	XRCC4_HUMAN	YEK7_SCHPO	YQ58_CAEEL
ZBT17_CHICK	ZBT17_HUMAN	ZBT20_HUMAN	ZBT24_MOUSE	ZF161_HUMAN	ZF64A_HUMAN
ZF64B_HUMAN	ZFP27_MOUSE	ZFP28_HUMAN	ZFP28_MOUSE	ZFP2_MOUSE	ZFP46_MOUSE
ZFP62_MOUSE	ZFP92_MOUSE	ZFP95_HUMAN	ZG16_XENLA	ZG48_XENLA	ZN124_HUMAN
ZN12_MICSA	ZN134_HUMAN	ZN137_HUMAN	ZN138_HUMAN	ZN160_HUMAN	ZN167_HUMAN
ZN174_HUMAN	ZN177_HUMAN	ZN197_HUMAN	ZN200_HUMAN	ZN202_HUMAN	ZN205_HUMAN
ZN206_HUMAN	ZN213_HUMAN	ZN228_HUMAN	ZN229_HUMAN	ZN236_HUMAN	ZN239_HUMAN
ZN253_HUMAN	ZN254_HUMAN	ZN268_HUMAN	ZN271_HUMAN	ZN272_HUMAN	ZN282_HUMAN
ZN287_MOUSE	ZN324_HUMAN	ZN333_HUMAN	ZN336_HUMAN	ZN343_HUMAN	ZN345_HUMAN
ZN350_HUMAN	ZN366_HUMAN	ZN37A_HUMAN	ZN393_HUMAN	ZN398_HUMAN	ZN434_HUMAN
ZN440_HUMAN	ZN444_HUMAN	ZN490_HUMAN	ZN497_HUMAN	ZN509_HUMAN	ZN513_HUMAN
ZN525_HUMAN	ZN575_HUMAN	ZN575_MACFA	ZN597_HUMAN	ZN621_HUMAN	ZN624_HUMAN
ZN660_HUMAN	ZN75A_HUMAN	ZNF16_HUMAN	ZNF22_MOUSE	ZNF31_HUMAN	ZNF35_HUMAN
ZNF35_MOUSE	ZNF36_HUMAN	ZNF42_HUMAN	ZNF69_HUMAN	ZNF70_HUMAN	ZNF71_HUMAN
ZNF74_HUMAN	ZNF75_HUMAN	ZNF77_HUMAN	ZNF79_HUMAN	ZNF80_CERAE	ZNF91_HUMAN
ZNF96_HUMAN	ZNFEB_HUMAN	ZO22_XENLA	ZO29_XENLA	ZO2_XENLA	ZO71_XENLA
RL10A_MOUSE	RL13_SACEX	RL14_TRYCO	RL18A_CAEEL	RL28_HUMAN	RL31_HUMAN
RL32P_MOUSE	RR14_EUGGR	RS6_SPOFR	RS7_SCHPO	ARP6_ASHGO	CND1_XENLA
GID8_YEAST	IMA2_ORYSA	IRF3_MOUSE	ISY1_CANGA	PSB2_YEAST	PSB4_ARATH
SLBP_HUMAN	TCL2_CAEEL				

B.

MYH11_RAT	MYH6_RABIT	MYH7_RABIT	MYSC_CHICK	MYSP_CAEEL	MYSP_DERFA
MYSP_OPIFE	MYSU_RABIT	MYS_PODCA	ADH1_ENTHI	AK1E1_MOUSE	DCTN4_MOUSE
ENOA_BOVIN	ENT3_YEAST	FABPE_HUMAN	GOLP3_HUMAN	GSTF1_ARATH	HMCS1_HUMAN
HOME2_HUMAN	HOOK3_HUMAN	INVO_GALCR	INVS_BRARE	K1C20_HUMAN	MANA_YEAST
MYPT1_RAT	MYPT2_MOUSE	RACH_DICDI	SCE3_SCHPO	ST1A1_HUMAN	VIAF1_DROME
VPS36_MOUSE	XFIN_XENLA	YPT3_SCHPO			

C.

AAT3_ARATH	ADAS_DROME	ADXH_DROME	AKA7A_HUMAN	AL1B2_BOVIN	C11B2_MESAU
CADH9_MOUSE	CBR2_MOUSE	CLIC5_BOVIN	COQ3_SCHPO	CTRA_GADMO	CYLC2_BOVIN
CYLC2_HUMAN	DECR2_BRARE	ENP1_TORCA	FETUA_PIG	FUT2_MOUSE	GIC1_YEAST
HBLD2_HUMAN	MYS2_DICDI	NACH_DROVI	NFU1_YEAST	NIDM_BOVIN	ODBA_RAT
OTC_RANCA	PA2B_TRIMU	PSAF_FLATR	RER2_YEAST	RK22_MAIZE	RK2_PINKO
RM43_BOVIN	RM43_HUMAN	RPOC1_GUITH	RR3_ARATH	SODM_BOVIN	SODM_CALJA
SPYA_CALJA	STAD_RICCO	TIM17_SCHPO	TIM23_BRARE	TRBP2_ARATH	YCF3_PSINU

2.2 Global score by analysis of whole amino acid sequences

We hypothesized in this work that the signals for the translocation of proteins to the nucleus are spread throughout a much wider region compared to normal sequence motifs, as suggested from the long periodicity of charges in PCP28. Two types of scores were developed to discriminate nuclear PCP28: the global score obtained by analysis of entire amino acid sequences, and the local score

obtained by analysis of fragments of 31 residues around the cluster of positively charged residues. For developing a global score of PCP28, we calculated the average values of various physicochemical parameters of entire sequences of two classes of PCP28: the nuclear and cytoplasmic PCP28:

$$\langle X \rangle_p = \left[\sum_{m=1}^{L} X_p(m) \right] \Big/ L \qquad (2)$$

where $X_p(m)$ is the value of the *p*-th property at the *m*-th residue in a sequence, and L is the length of the protein. Here, the selected properties were the densities of negatively charged residues (Asp/Glu), aromatic residues (Trp/Tyr/Phe), proline, glycine, and cysteine, the amphiphilicity index[3], and the hydropathy index[4] of amino acids. In this case, *p* is a value from 1 to 7. When the average values of a parameter are significantly different between the two classes of PCP28, the parameter will contribute to the prediction of the nuclear PCP28. Therefore, we selected seven parameters based on the significance of the difference.

The parameter $\langle X \rangle_p$ was normalized by the following equation:

$$Z_p^{(\mathrm{global})} = \left(\langle X \rangle_p - \overline{\langle C \rangle_p} \right) \times \left(\overline{\langle N \rangle_p} - \overline{\langle C \rangle_p} \right) \qquad (3)$$

where p represents the number for seven parameters, and $\overline{\langle N \rangle_p}$ and $\overline{\langle C \rangle_p}$ are the *p*-th parameter averaged over all data of the nuclear PCP28 and over the cytoplasmic PCP28, respectively. The normalized parameters $Z_p^{(\mathrm{global})}$ (p=1~7) were used for the discrimination analysis.

Finally, the score $S^{(\mathrm{global})}$ was calculated using discrimination analysis technique, which is expressed by the linear combination of seven parameters $Z_p^{(\mathrm{global})}$:

$$S^{(\mathrm{global})} = \sum_{p=1}^{7} \left(a_p Z_p^{(\mathrm{global})} \right) \qquad (4)$$

2.3 Local score obtained using analysis of local sequences around charge symmetry peaks

Inspecting the charge distribution around the peak of positive charge at the nuclear localization signals, we found that the charge distribution appears symmetric. Therefore, we defined the charge symmetry score using the following equation[10]:

$$S_c(i, j) = \frac{1}{j} \sum_{k=1}^{j} \left\{ \overline{q(i+k)} + \overline{q(i-k)} \right\}^2 \qquad (5)$$

where $S_c(i, j)$ is the charge symmetry score at the sequence position of i and for a width of 2j+1, and q is the electric charge of amino acid residues (K,R,H = +1; D,E = -1). We scanned the width from 5 to 15 residues at a fixed position, and the maximum values of $S_c(i, j)$ at position i were used for the charge symmetry score.

Figure 1 shows the plots of net charge and charge symmetry (A) and the sequence regions used for the analysis of the local score (B). The upper graph shows the plot of the moving average of electric charges, while the lower graph is the corresponding plot of charge symmetry calculated using Eq. (5). A significant peak was observed at the center of the cluster of positively charged residues. Because the nuclear localization signals are positively charged, we used only charge symmetry peaks from positive charge clusters. To take the distribution of various properties into account, we chose five regions around the charge symmetry peak, as shown in Figure 1B: a region of eleven residues including the charge symmetry peak and two regions of five residues on both sides of the central region. Therefore, five average values for each property were calculated.

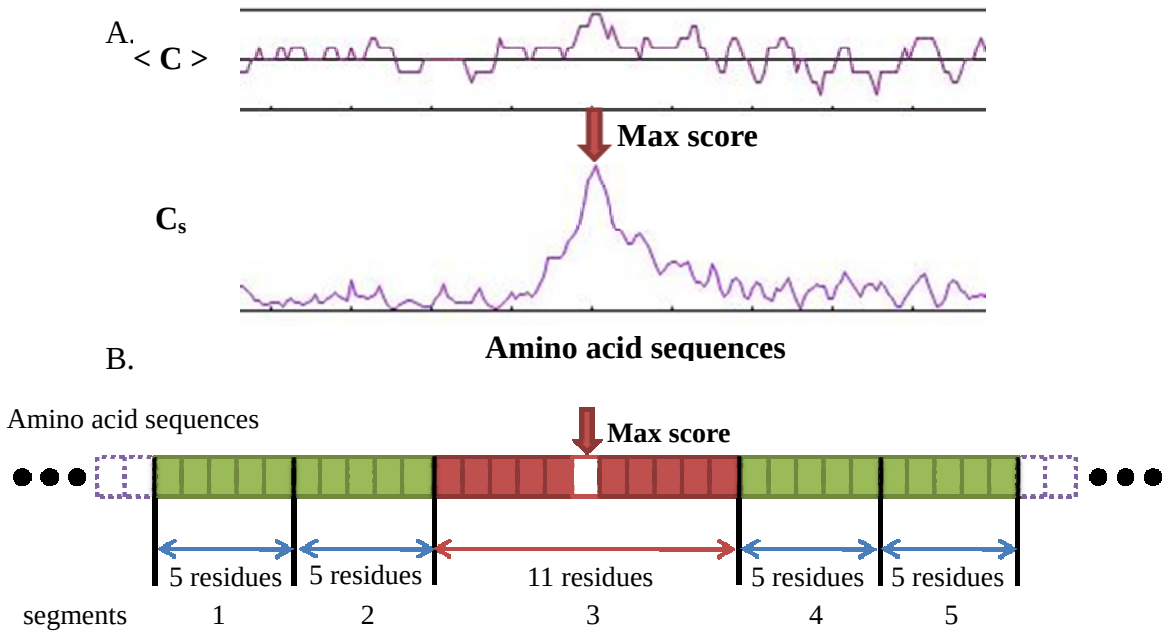

Figure 1. Moving average of electric charges (<C>), charge symmetry score (C_s) (A), and definition of the five segments (B).
<C> was calculated using equation (6); C_s was calculated using equation (5)

Calculating the local predictive score is similar to calculating the global score. However, the local calculation is complicated because the region involving 31 residues is divided into five smaller regions. First, the moving averages of the *p*-th property $\langle X_p(i) \rangle_7$ were calculated using:

$$\langle X_p(i) \rangle_7 = \left[\sum_{j=i-3}^{i+3} X_p(j) \right] \Big/ 7 \qquad (6)$$

The properties for analysis of the local score included the density of positively charged residues (Lys/Arg), negatively charged residues (Asp/Glu), small polar residues (Ser/Thr/Asn/Asp), aromatic residues (Trp/Tyr/Phe), and cysteine, and the amphiphilicity index of amino acids. Then, the double average for five regions (Figure 1B) around the charge symmetry peak was calculated using:

$$\left\langle\left\langle X_p \right\rangle\right\rangle_k = \sum_{i \in segment(k)} \left\langle X_p(i) \right\rangle_7 \Big/ l \tag{7}$$

where $\left\langle\left\langle X_p \right\rangle\right\rangle_k$ is the double average of the *p*-th property in the *k*-th region, and *l* is the 11 residues of the center segment and 5 residues from other regions.

The double average $\left\langle\left\langle X_p \right\rangle\right\rangle_k$ is normalized by the difference between the nuclear and cytoplasmic proteins by Eq. (8), leading to the *p*-th parameter $Z_p^{(\text{local})}$.

$$Z_p^{(\text{local})} = \sum_{k=1}^{5} \left(\left\langle\left\langle X_p \right\rangle\right\rangle_k - \overline{\left\langle\left\langle C_p \right\rangle\right\rangle_k} \right) \times \left(\overline{\left\langle\left\langle N_p \right\rangle\right\rangle_k} - \overline{\left\langle\left\langle C_p \right\rangle\right\rangle_k} \right) \tag{8}$$

where $\overline{\left\langle\left\langle N_p \right\rangle\right\rangle_k}$ and $\overline{\left\langle\left\langle C_p \right\rangle\right\rangle_k}$ are the average values of $\left\langle\left\langle X \right\rangle\right\rangle_k$ for all nuclear and all cytoplasm data, respectively. Finally, the local score $S^{(\text{local})}$ was determined by discrimination analysis with seven parameters–the six parameters $Z_p^{(\text{local})}$ (p=1~6) and charge symmetry value C_{peak} at the peak.

$$S^{(\text{local})} = \sum_{p=1}^{6} \left(b_p Z_p^{(\text{local})} \right) + b_7 C_{\text{peak}} \tag{9}$$

The coefficients b_1~b_7 in the score $S^{(\text{local})}$ were determined by discrimination analysis of the nuclear and cytoplasmic PCP28.

2.4 Discrimination score as a linear combination of the global and local scores

The unified score S was determined as a linear combination of the two scores, $S^{(\text{global})}$ and $S^{(\text{local})}$, to give the best classification between nuclear and cytoplasmic PCP28.

$$S = c_1 S^{(\text{global})} + c_2 S^{(\text{local})} \tag{10}$$

The global and local parameters are independent of each other; therefore, the contribution of the two scores, $S^{(\text{global})}$ and $S^{(\text{local})}$, to the unified score S will provide the mechanism of the nuclear translocation.

3. Results

We selected PCP28 from proteins of eukaryotes deposited in Swiss-Prot and classified them in terms of subcellular localization provided by the features table of Swiss-Prot. Calculation of the autocorrelation function of the charge distribution of all selected PCP28, shown in Figure 2, revealed significant peaks at multiples of 28 residues. These peaks were observable even at approximately 300 residues. In addition, more than half of the PCP28 were localized in the nucleus, as shown in Table 1, even though completely different classes of proteins, such as motor proteins, also belonged to the PCP28. As seen in Table l, the selection of PCP28 from amino acid sequences

corresponds to the screening of nuclear proteins with an accuracy of about 70%. Therefore, the discrimination of nuclear PCP28 from other types of PCP28 leads to a very accurate prediction of nuclear proteins.

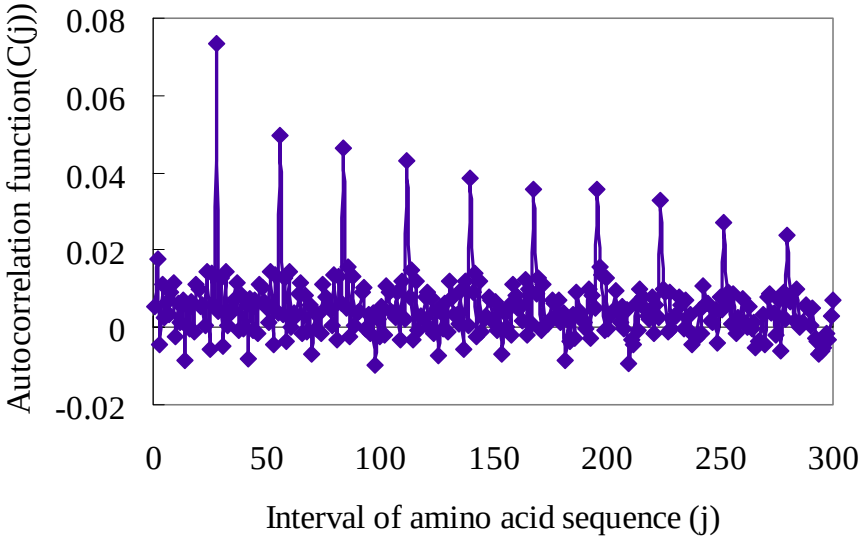

Figure 2. Autocorrelation function of PCP28 of eukaryotic proteins extracted from Swiss-Prot database rel.48.7.

3.1 Differences in physicochemical properties of nuclear and cytoplasmic PCP28

Considerable differences exist in the average physicochemical properties of whole amino acid sequences between nuclear and cytoplasmic PCP28 as well as in the distribution of properties around the clusters of positively charged residues which are characteristic of the nuclear localization signals.

Figure 3A is a comparison of seven properties of whole sequences between nuclear and cytoplasmic PCP28. This diagram demonstrates that entire sequences have several systematic property differences. Local structures are usually considered essential for molecular recognition and the translocation of proteins to various subcellular compartments. However, for PCP28, the properties of entire amino acid sequences appear to contribute to nuclear localizations. Seven parameters were considerably different between the nuclear and cytoplasmic PCP28: hydrophobicity index; amphiphilicity index; and the densities of negatively charged residues, aromatic residues, proline, glycine, and cysteine. Some of the properties (*e.g.*, the densities of proline and glycine and the amphiphilicity index) are signals of the disordered regions[11, 12], suggesting that the nuclear PCP28 contain a large fraction of disordered regions throughout the sequences.

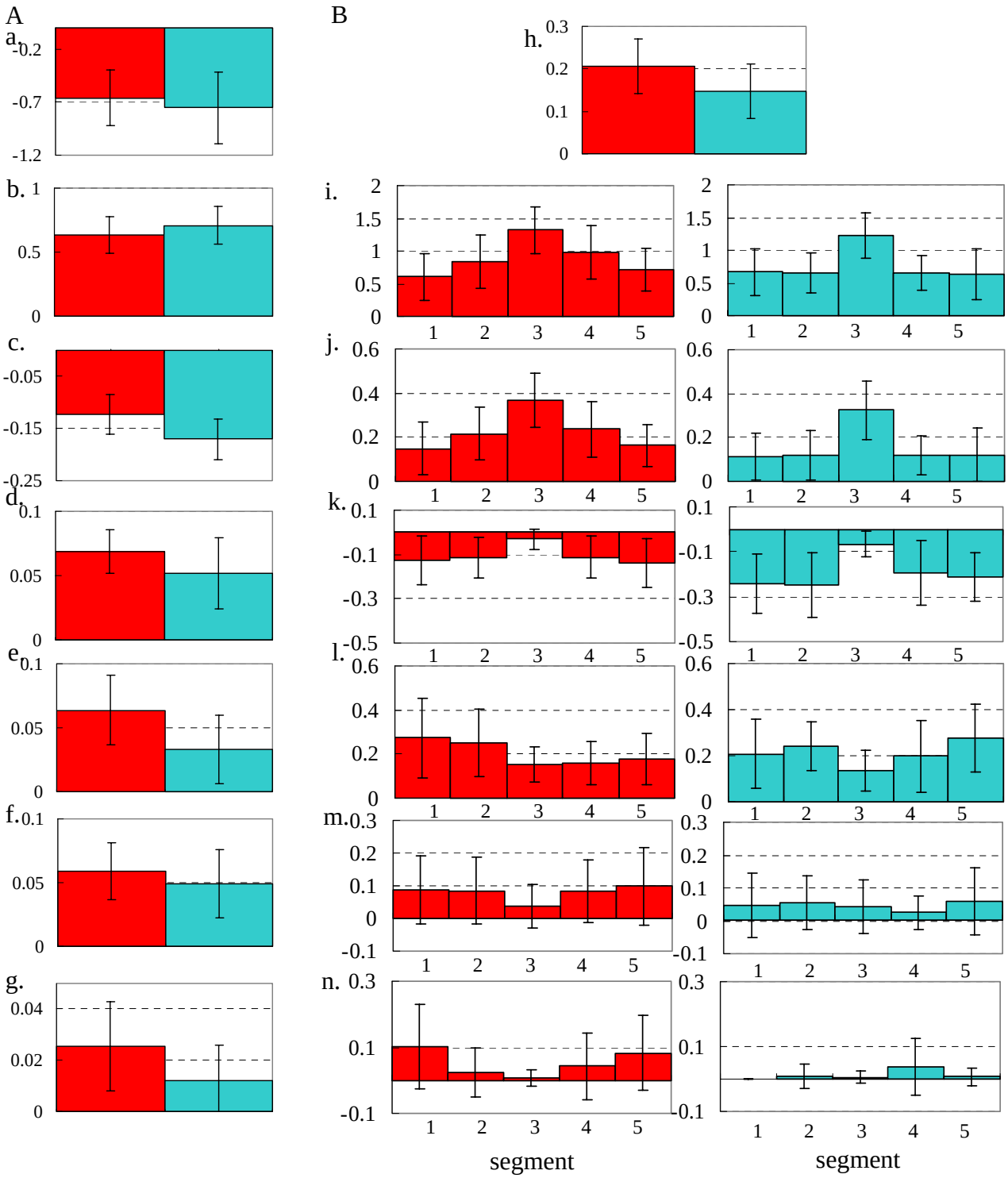

Figure 3. The distribution of physicochemical properties for PCP28 in the nucleus and cytoplasm.

A) Comparison of seven properties of entire sequences: (a) hydrophobicity index, (b) amphiphilicity index, (c) negatively charged residues, (d) aromatic residues, (e) proline, (f) glycine, and (g) cysteine.

B) Comparison of seven properties of local sequences around charge symmetry peaks: (h) charge symmetry peak values, (i) amphiphilicity index, (j) positively charged residues, (k) negatively charged residues, (l) small polar residues, (m) aromatic residues, and (n) cysteine.

[Red represents nuclear PCP28 and blue represents cytoplasmic PCP28.]

For analysis of local properties of nuclear PCP28, attention was focused on the clusters of electric charges because nuclear localization signals usually have large positive charges. The charge symmetry index of Eq. (5) has a peak at the center of a cluster of the net charge and becomes zero at the interface between the positive and negative net charges in sequences. Therefore, peaks of the charge symmetry index were selected as candidates of nuclear localization signals. Figure 3B compares the distribution of various properties around the true nuclear localization signals in nuclear PCP28 and the false candidates in cytoplasmic PCP28. In addition to the peak values of the charge symmetry index, six properties were analyzed, dividing the 31 residues around the charge symmetry peaks into five regions: the amphiphilicity index, densities of positively and negatively charged residues, small polar residues, aromatic residues, and cysteine. Figure 3B compares the differences in distribution of the seven properties of the nuclear and cytoplasmic PCP28. The charge symmetry peak is higher at the true nuclear localization signals than at the false ones. Positive charges are abundant at neighboring regions of the nuclear localization signals, while the distribution of negative charges shows the reverse tendency, indicating that electrostatic repulsion dominates even in local segments of the nuclear localization signals. The significant difference in the density of cysteine is probably due to the zinc finger motif in the nuclear PCP28. It should be pointed out here that the nuclear localization signals, which have been confirmed experimentally, are very few and the peaks of the charge symmetry may not be the real nuclear localization signals. We have used the charge symmetry peaks as candidates of the binding sites with the transporting machinery and examined the accuracy of the discrimination system between the nuclear and the cytosolic PCP28.

3.2 Discrimination analyses using global and local properties

Discrimination analysis was conducted using Eq. (4) for the global parameters in Figure 3A, and Eq. (9) for the local parameters in Figure 3B, leading to the discrimination scores, $S^{(\mathrm{global})}$ and $S^{(\mathrm{local})}$, respectively. The final discrimination score S was provided by Eq. (10) as a linear combination of the two scores, $S^{(\mathrm{global})}$ and $S^{(\mathrm{local})}$. Figures 4A-4C show the results of the three discrimination analyses for the training dataset. Figures 4A and 4B show the population of PCP28 as functions of $S^{(\mathrm{global})}$ and $S^{(\mathrm{local})}$, respectively. Discrimination by the global and the local score was similarly good, and a simple estimation of the accuracy of the discrimination using the single scores led to a value between 80% and 90%. Then, the two scores were combined for better accuracy. Figure 4C shows the result of the prediction using the unified score for the training dataset. The performance of this method was better than 90%. The contributions of various parameters in the discrimination analyses are shown in Table 3. The contribution of the parameters can be estimated by multiplying the weights and the differences of average values.

Figure 5 shows the results of the discrimination analysis for the test dataset, using all PCP28 from Swiss-Prot. In this dataset, many data did not have annotations about subcellular localization. Thus, the data from the nuclear proteins, cytoplasmic proteins, and proteins without subcellular localization information were plotted in separate graphs. The sensitivity and specificity of the prediction between the nuclear and cytoplasmic PCP28 were 92% and 88%, respectively.

We also carried out a cross validation test, using the 30 positive and 30 negative data points for the training dataset and testing the system by the remaining data points. The average values for the accuracy of the sensitivity and the specificity of ten cross validation tests were 88% and 84%, respectively. These values are only 4% less than the results obtained by the test dataset which included all the training data.

Table 2. Contribution to discrimination scores.

Contribution to the: (A) global score, (B) local score, and (C) unified score. The contribution is calculated by multiplying the weight and difference in Z_p

A.

Property	Hydrophobicity index	Amphiphilicity index	Densities				
			Neg	Aromatic	PRO	GLY	CYS
Weight	1.54	-7.96	10.77	11.76	26.37	-34.2	30.98
Difference of $\overline{Z_p^{(global)}}$	0.034	0.039	0.060	0.017	0.035	0.004	0.010
Contribution	0.05	-0.31	0.64	0.20	0.92	-0.15	0.32

B.

Properties	Amphiphilicity index	Densities					Charge Symmetry index
		Pos	Neg	STDN	Aromatic	CYS	
Weight	0.24	2.4	0.66	3.77	3.84	4.14	6.3
Difference of $\overline{Z_p^{(local)}}$	0.387	0.233	0.469	0.129	0.073	0.137	0.058
Contribution	0.09	0.56	0.31	0.49	0.28	0.57	0.37

C.

Parameter	Global score	Local score
Weight	0.36	0.84
Difference of average	1.70	2.66
Contribution	0.61	2.24

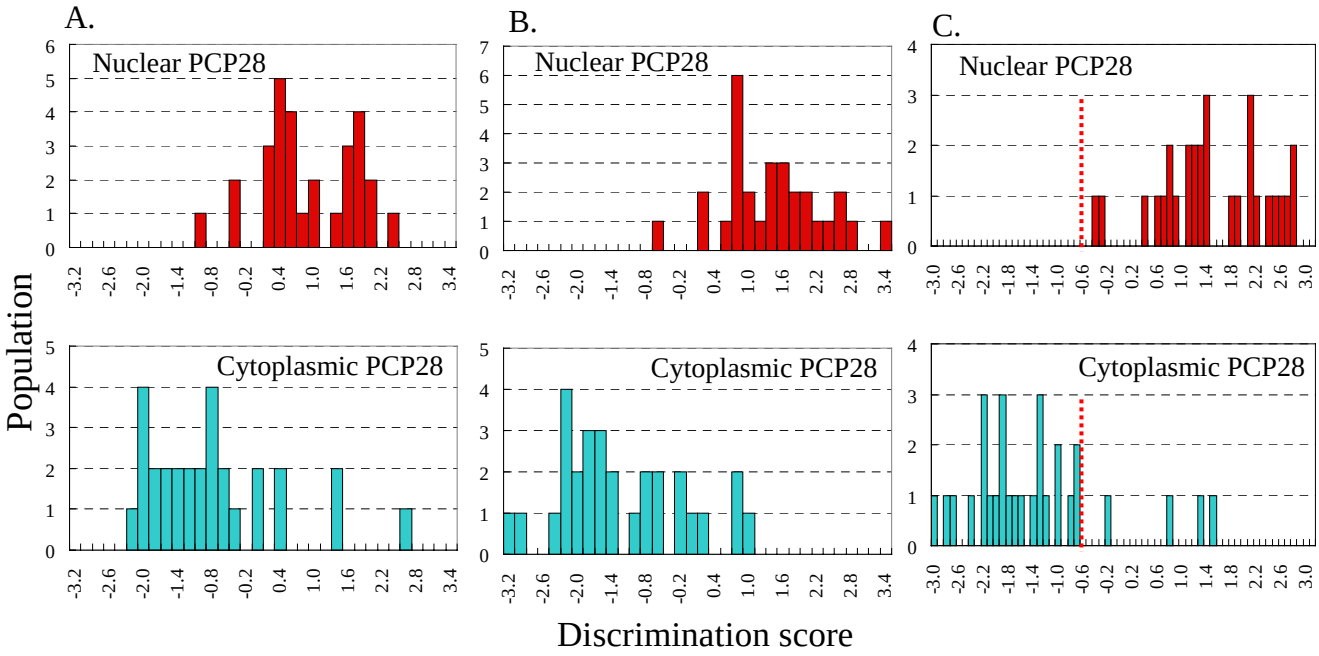

Figure 4. Discrimination between PCP28 localized in nucleus (red) and in cytoplasm (blue).

Results of the discrimination analyses are shown for the: (A) global score, (B) local score, and (C) unified score.

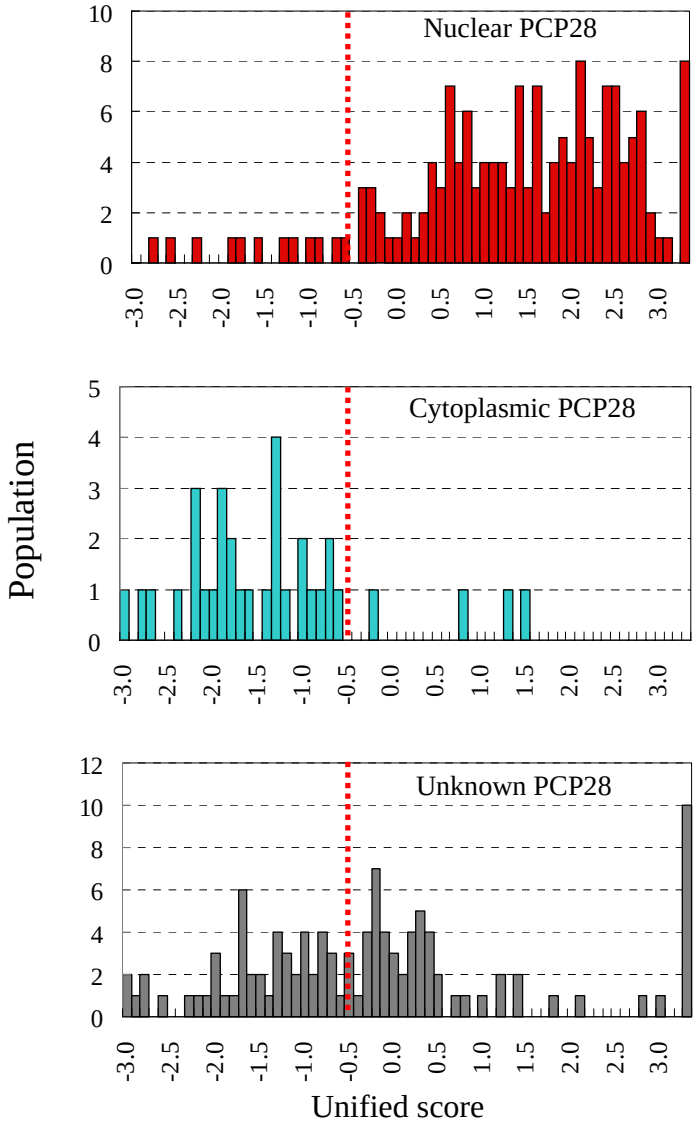

Figure 5. Discrimination analysis of test dataset of all PCP28 from Swiss-Prot.
PCP28 localized in the nucleus (red), cytoplasm (blue), and unknown localization (gray) are shown. Threshold of the unified score is -0.6 for the best discrimination.

4. Discussion

Several interesting conclusions are suggested by the present results.

1) A good system for the detection of the differences in the physicochemical properties between the nuclear and cytoplasmic PCP28 was obtained by using segments, much longer than the usual sequence motifs.
2) Classification of nuclear PCP28 was successful when seven parameters were used for the global and the local scores. The success of the discrimination suggests that difficulty in understanding subcellular localization is partly due to the contribution of many factors to the phenomenon.
3) The present system is completely different from other prediction methods, such as sequence

homology, sequence motifs and various information science methods, such that, the average values of physicochemical properties were used instead of the characters themselves. Moreover, this approach may be generalized to other phenomena as well. The prediction system SOSUI for membrane proteins was developed by the same approach.

A sequence motif is usually considered much shorter than 28 residues. However, the autocorrelation function of the electric charge sequences in PCP28 shows very sharp 28-residue periodicity over 300 residues, as shown in Figure 2. Furthermore, the global score using the average values of various properties of entire sequences contributed to the final discrimination function. These suggest that a sequence that determines the translocation of a protein into the nucleus can be as long as 100 residues. The size of the segments that determines molecular recognition is closely related to the ambiguity of the sequence motif because the average properties were used in our approach. For discrimination analysis of the nuclear PCP28, the average properties of 31 residues were used for calculating the local score, while the average of the entire amino acid sequence was used for the global score. This approach was successful in developing a predictor with sensitivity and specificity greater than 80%, indicating that sequence information within a resolution of one residue is not necessary for molecular recognition of the nuclear proteins. In other words, information about clusters of amino acids is important for the translocation phenomenon. A method for accurate discrimination of the nuclear PCP28 from other soluble PCP28 will be very useful for understanding the mechanism of the translocation of proteins into the nucleus.

The contribution of the parameters to the discrimination scores provides information about the mechanism of the nuclear localization of proteins. Table 2 shows the parameters used for such analyses. Note that the contribution of the global score to the final discrimination score is as large as 20%. This may suggest a hypothesis such that the nuclear localization of proteins is determined not only by local molecular recognition but also by properties of the entire sequence, probably related to the flexibility of the proteins, as inferred from the large contribution of proline to the global score. However, many properties in fact contribute to the discrimination scores, such as $S^{(\mathrm{global})}$ and $S^{(\mathrm{local})}$; suggesting simple sequence motifs cannot discriminate the nuclear proteins because of the contribution of various physicochemical properties of sequences. This concept has not yet been confirmed experimentally but it is an interesting problem that needs to be looked further into in the future.

Structures and functions of proteins are thought to develop through physical interactions, in principle. However, the type of interaction essential for protein processes has not been elucidated. In this work, seven properties of amino acid sequences were used for developing a high performance predictive system of nuclear proteins. This suggests that no dominant interactions exist, but that many properties contribute to protein processes. Furthermore, selection of the appropriate amino acid sequence properties allows prediction with high accuracy. The first example of a predictive system like this was the membrane protein predictor SOSUI, which used two physicochemical indices, the hydrophobicity and the amphiphilicity indices. The average values for several residues were enough to develop an accurate system. Discrimination of nuclear PCP28 is more complicated than that of membrane proteins, but similar principles apply to the two types of proteins, and the accuracy of the predictive systems is also similar.

Finally, this approach to development of a protein predictive system is closely tied to the mechanism of protein processes. For example, transmembrane helices were predicted by two parameters, the average hydrophobicity at the center of the helices and the average amphiphilicity at the helix ends. This algorithm was based on the stabilization of a transmembrane helix through interaction with the lipid bilayer, which is hydrophobic at the center and amphiphilic at the hydrocarbon/water interface[5]. Similarly, the parameters for discrimination of nuclear PCP28 are probably closely related to the mechanism of nuclear localization. Although details of the nuclear

localization mechanism are not yet known, this work indicates that the structural flexibility and electric repulsion within proteins are important for the translocation process.

Ongoing analysis of entire amino acid sequences from biological genomes suggests that vertebrate genomes code more PCP28 than other biological genomes. In the case of the analysis of all amino acid sequences from the human genome, the nuclear PCP28 represents about 2% of the total percentage of the human proteome. The results of this continuing work will be reported at a future date.

This work was partially supported by the 21st Century COE of Frontiers of Computational Science at Nagoya University.

References

[1] Hirokawa T., Boon-Chieng S. and Mitaku S., SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, **14**, 378-379(1998).

[2] Tsuji T. and Mitaku S., Features of transmembrane helices useful for membrane protein prediction, *CBIJ*, **4**, 110-120(2004).

[3] Mitaku S., Hirokawa T. and Tsuji T., Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces, *Bioinformatics*, **18**, 608-616(2002).

[4] Kyte J. and Doolittle R.F., A Simple Method for Displaying the Hydropathic Character of a Protein, *J. Mol. Biol.*, **157**, 105-132(1982).

[5] Hayashibe T., Hirokawa T. and Mitaku S., Novel Index of Polar Amino Acids Characterizing End Region of Transmembrane Helices, *Genome Informatics*, **11** 416-417(2000).

[6] Ke R., Sakiyama N., Sawada R., Sonoyama M. and Mitaku S., Human Genome Encodes Many Proteins with Charge Periodicity of 28 Residues, *Jpn. J. Appl. Phys.*, **46**, 6083-6086 (2007).

[7] Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.*, **31**, 365-370(2003).

[8] Kalderon D., Richardson W.D., Markham A.F. and Smith AE., Sequence requirements for nuclear location of simian virus 40 large-T antigen, *Nature*, **311**, 33-38(1984).

[9] Robbins J., Dilworth S.M., Laskey R.A. and Dingwall C., Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence, *Cell*, **64**, 615-623(1991).

[10] Imai K., Asakawa N., Tsuji T., Sonoyama M. and Mitaku S., Secondary structure breakers and hairpin structures in myoglobin and hemoglobin, *CBIJ*, **5**, 65-77(2005).

[11] Dunker A.K., Lawson J.D., Brown C.J., Williams R.M., Romero P, Oh J.S., Oldfield C.J., Campen A.M., Ratliff C.M., Hipps K.W., Ausio J, Nissen M.S., Reeves R., Kang C., Kissinger C.R., Bailey R.W., Griswold M.D., Chiu W., Garner E.C. and Obradovic Z.A., Intrinsically disordered protein, *J. Mol. Graph. Model*, **19**, 26-59(2001).

[12] Imai K. and Mitaku S., Mechanisms of secondary structure breakers in soluble proteins, *BIOPHYSICS*, **1**, 55-65(2005).