CROATICA CHEMICA ACTA CCACAA **75** (3) 805–816 (2002)

ISSN-0011-1643
CCA-2833

Review

Intramolecular Autoregulatory Sequences in Proteins*

Bostjan Kobe

Department of Biochemistry and Molecular Biology, and Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia
(E-mail: b.kobe@mailbox.uq.edu.au)

Received November 21, 2001; revised April 2, 2002; accepted April 19, 2002

The worldwide structural genomics initiative promises, within the next decade, to provide us with the three-dimensional structures of most representative proteins. However, the initiative does not in the initial stages specifically address the structural basis of protein-protein interactions and protein regulation, both of which are crucial for understanding the cellular function of proteins. This report reviews the studies of the structural basis of protein-protein interactions and regulation in several biologically important cellular processes carried out in our laboratory. The focus will be on the recognition of nuclear localization sequences (NLSs) by the nuclear import factor importin-α, and its regulation. Regulatory interactions in retroviral envelope proteins and the enzyme phenylalanine hydroxylase (PAH) will be discussed also. All three biological systems share a regulatory mechanism termed active site-directed or intrasteric regulation.

Key words: active site-directed/intrasteric regulation, importin-α/karyopherin-α, membrane fusion, nuclear localization signal, phenylalanine hydroxylase, protein-protein interactions, protein structure, viral envelope proteins, X-ray crystallography.

INTRODUCTION

The function of a protein molecule depends on its three-dimensional structure. It is becoming increasingly clear that structure does not only explain the known functional properties of a protein, but that much can be in-

* Based upon the plenary lecture presented at the 10th Croatian-Slovenian Crystallographic Meeting, Lovran, Croatia, June 21–24, 2001.

ferred about the function of a protein (with yet unknown function) from knowing its structure. This observation forms the foundation of a new field of structural biology termed »structural genomics«, a systematic program of high-throughput structure determination aimed at determining the structure of a representative protein from every protein family.[1] Several structural genomics initiatives have started worldwide and are already producing valuable data.

Most proteins in the cell, however, do not function on their own; they interact with other proteins and other molecules such as DNA, RNA and small molecule ligands. Furthermore, many proteins are regulated in complex ways, and enzymes catalyze complex chemical reactions. Studies of these aspects of protein function are poorly covered by structural genomics. Our laboratory has been especially interested in understanding the structural basis of protein-protein interactions and protein regulation, using primarily X-ray crystallography, and combining this method with biophysical, biochemical and molecular biology techniques that yield quantitative information on the interactions and link structural properties to functional aspects. Our integrated approach to studying protein-protein interactions will be illustrated by our work on the process of protein import into the cell nucleus, the metabolic enzyme phenylalanine hydroxylase, and the mechanism of membrane fusion catalyzed by retroviral fusion proteins.

THE ROLE OF IMPORTIN-α IN THE PROCESS OF NUCLEAR IMPORT

Unlike bacterial cells, all eukaryotic cells contain a cell nucleus. Many important processes, including transcription, occur in the nucleus. However, proteins are synthesized in the cytoplasm, and proteins that function in the nucleus need to be transported there. The nucleus is separated from the cytoplasm by a double membrane termed the nuclear envelope. The transport occurs through nuclear pore complexes (NPCs), large complex structures that penetrate the double lipid layer of the nuclear envelope. Most macromolecules require an active, signal-mediated transport process. The first and best characterized nuclear targeting signals are the classical nuclear localization sequences (NLSs) that contain one or more clusters of basic amino acids.[2] The NLSs do not conform to a specific consensus sequence, and fall into two distinct classes termed monopartite NLSs, containing a single cluster of basic amino acids, and bipartite NLSs, containing two basic clusters.

Despite the variability, the classical NLSs are recognized by the same receptor protein termed importin or karyopherin, a heterodimer of α and β

subunits.[3] Importin-α (Impα) contains the NLS-binding site and importin-β (Impβ) is responsible for the translocation of the importin-substrate complex through the NPC (Figure 1). The transfer through the pore is facilitated by the proteins Ran (Ras-related nuclear protein) and nuclear transport factor-2 (NTF2). Once inside the nucleus, Impβ binds to Ran-GTP, which causes the dissociation of the import complex and the autoinhibition of Impα; the importin subunits return to the cytoplasm separately without the import cargo. The directionality of nuclear import is thought to be conferred by an asymmetric distribution of the GTP- and GDP-bound forms of Ran between the cytoplasm and the nucleus. This distribution is in turn controlled by various Ran-binding regulatory proteins.

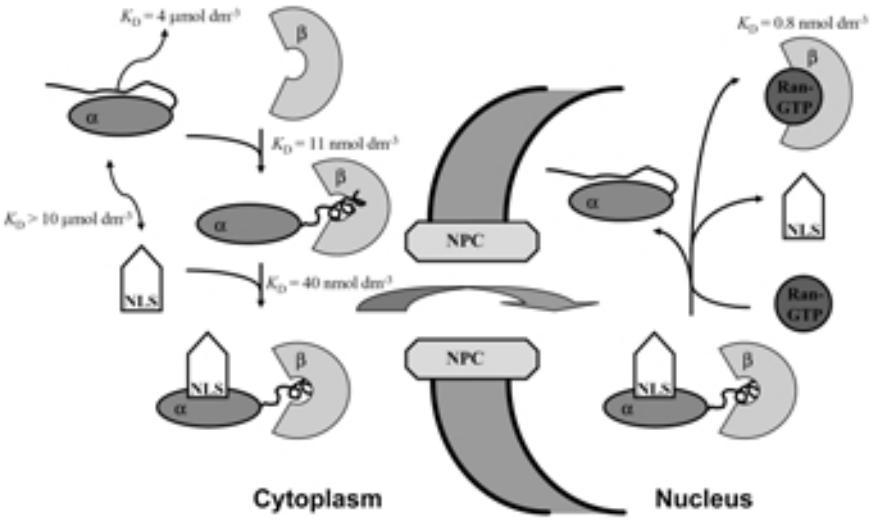

Figure 1. Schematic diagram of the NLS-dependent nuclear import pathway. Importin-α, oval grey object »α«; importin-β, light-grey object »β«; NLS-containing cargo protein, white pentagonal object »NLS«; Ran-GTP, round dark-grey object »Ran GTP«. For simplicity, other factors involved in the pathway such as NTF2, the nuclear export receptor for importin-α and Ran-binding proteins have been omitted from the diagram. Dissociation constants for the different binding events, based on our surface plasmon resonance results,[7] are shown.

To understand the variety of interactions formed by Impα during a cycle of protein import, we set out to determine the crystal structures of Impα and its complexes with peptides corresponding to NLSs, and characterize the interactions using surface plasmon resonance on a BIAcore biosensor. The first structure we determined was of the intact, recombinantly expressed mouse Impα.[4] The structure of a truncated yeast Impα published some-

what earlier [5] provided the opportunity to compare the two proteins, and revealed an interesting function for the N-terminal region of the protein. It was previously established that Impα consists of two functional domains, a short N-terminal domain involved in Impβ binding (»IBB« domain), and a large C-terminal domain involved in NLS binding. While the IBB domain was not present in the yeast Impα used for structure determination, our structure revealed that the IBB domain interacted with the NLS binding site, autoinhibiting the protein (Figure 2).

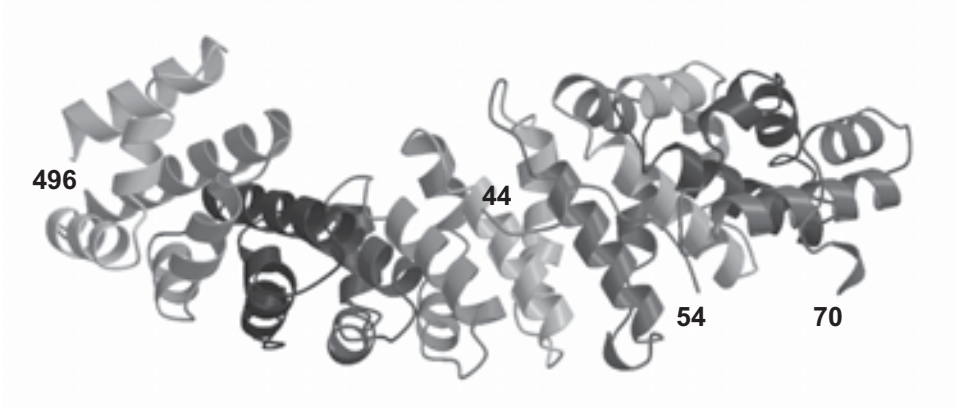

Figure 2. Structure of mouse importin-α.[4] Ribbon diagram of importin-α is shown (Figures 2, 4 and 5 are drawn with programs MOLSCRIPT and RASTER3D).[21,22] The N-terminal autoinhibitory segment is the region between residues 44 and 54.

Our structural observation suggested that Impα uses a so-called intrasteric or active site-directed mechanism of regulation.[6] In the nucleus, Impα exists as a monomeric protein and binding to nuclear proteins containing NLSs is not desired; the autoinhibitory IBB domain therefore prevents the binding of various nuclear proteins. Once transported to the cytoplasm, however, Impβ binds to the IBB domain, removing it from the NLS-binding site and activating Impα. In the cytoplasm, the Impα-Impβ complex can therefore collect NLS-containing proteins destined for the nucleus and transport them there. Once the trimeric transport complex reaches the nucleus, however, the protein Ran-GTP binds to Impβ and displaces Impα. Impα can therefore release its cargo, and return to the cytoplasm without Impβ or the cargo protein for the next cycle of import.

The model of regulation of nuclear import suggested based on our structure is supported by our affinity measurements using the BIAcore biosensor, and analogous experiments using different techniques by other groups (Ref.

7, and references therein). There appears to be an at least 250-fold increase in affinity for NLS binding by Impα if Impβ is present. We are further testing the regulatory mechanism using site-directed mutagenesis and nuclear import assays.

It used to be puzzling how a single receptor protein, Impα, can bind a diverse set of NLSs including monopartite ones such as the NLS from the simian virus 40 large T-antigen (T-Ag; PKKRKV, single letter amino acid code), and bipartite ones such as the NLS from nucleoplasmin (KRPAATKKAGQAKKK). Furthermore, either group of NLSs has very diverse sequences, with no obvious consensus sequence. Our structures of complexes of mouse Impα with peptides corresponding to NLSs,[8] and similar studies on the yeast protein,[9] have now resolved this puzzle. The two clusters of basic residues in bipartite NLSs bind to two distinct regions on the surface of Impα, using elec-

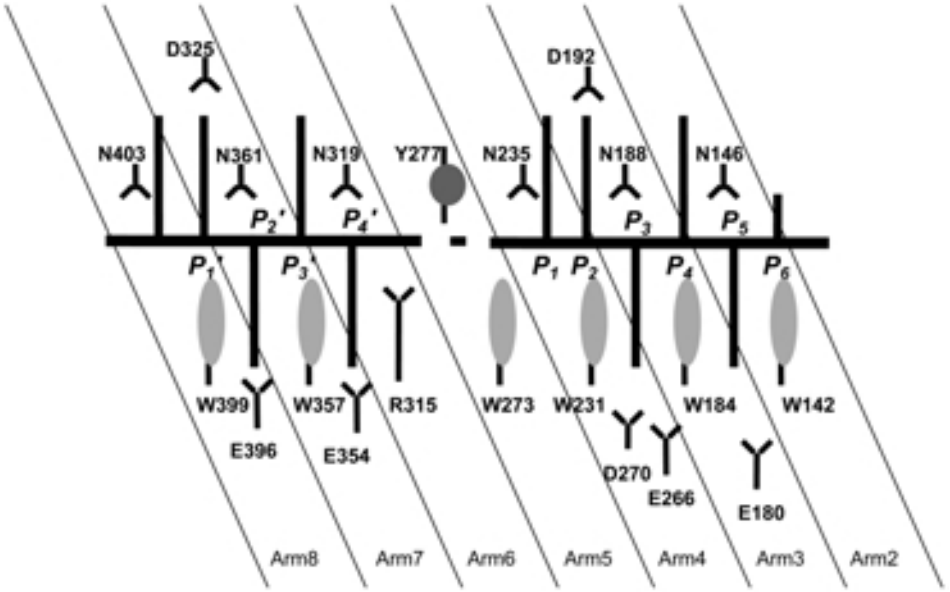

Figure 3. Schematic diagram of the interactions of NLS peptides with importin-α. The NLS backbone is shown as a thick horizontal line, with the side chains shown as perpendicular lines radiating from it. The NLS-binding domain of importin-α has a repetitive amino acid sequence pattern termed armadillo repeats, each repeat corresponding to a structural unit comprising three α-helices in the crystal structure;[4] individual armadillo repeats are separated by tilted lines in the Figure. Some importin-α side chains interacting with the NLS peptides are indicated: the invariant asparagines, the invariant tryptophans, and some nearby negatively charged residues. Y277 and R315 that interrupt the regular asparagine and tryptophan array are also shown.

trostatic, polar and hydrophobic interactions, while the linker sequence between the two clusters makes fewer favorable contacts and therefore does not need to be highly conserved. We have determined structures of complexes of Impα with peptides corresponding to several different bipartite NLSs and find that the linker sequence can form a diverse set of interactions, depending on its sequence and length (unpublished results). The basic cluster in monopartite NLSs can use either basic cluster binding site for binding, but the one used by the C-terminal region of the bipartite NLSs appears to be the high affinity site. The binding strategy used is extremely elegant, and explains the »promiscuous specificity« of NLS binding; individual side chain-binding pockets can often accommodate either a lysine or arginine residue, determining the specificity of binding, but the strongest interactions appear to be made by the main chain of the peptide. A schematic representation of the binding is shown in Figure 3.

Phosphorylation in the vicinity of NLSs sequences in some proteins provides yet another level of complexity and the opportunity to finely regulate nuclear import.[10] One system under complex control by phosphorylation is the T-Ag; phosphorylation N-terminal to the NLS increases the efficiency of nuclear import. Our preliminary studies on the structure of the complex of the phosphorylated and un-phosphorylated peptides corresponding to the relevant region of T-Ag with Impα reveals an intricate mechanism of regulation that will require further studies to be fully understood (unpublished results).

REGULATION OF PHENYLALANINE HYDROXYLASE

Phenylalanine hydroxylase (PAH) is a metabolic enzyme that converts phenylalanine to tyrosine using molecular oxygen, enzyme-bound iron, and a 6R-tetrahydrobiopterin (BH_4) cofactor.[11] It is a member of the aromatic amino acid hydroxylase family, together with tyrosine hydroxylase (TH) and tryptophan hydroxylase (TPH); TH and TPH are involved in the biosynthesis of the neurotransmitters, L-DOPA and serotonin, respectively. The aromatic amino acid hydroxylases share a similar enzyme mechanism and have a common three-domain structure consisting of an N-terminal regulatory domain, a catalytic domain and a C-terminal tetramerization domain; the highest sequence and structural similarity is found in the catalytic domain. Over 300 different mutations in the PAH gene have been found to be associated with the disease phenylketonuria (PKU).[12]

PAH, TH and TPH are regulated in different ways, and their regulatory domains are not highly conserved. PAH needs to be regulated very tightly, because it keeps under control the level of phenylalanine, an essential ami-

no acid, which is subject to large fluctuations due to dietary intake. On the one hand, the levels of PAH in the liver are such that if uncontrolled, the enzyme would rapidly deplete the phenylalanine stores; on the other hand, the metabolites of phenylalanine are toxic to the developing brain. The major regulatory mechanisms of PAH include activation by phenylalanine, inhibition by BH_4, and additional activation by phosphorylation.[11] Activation by the substrate phenylalanine is considered the major regulatory event, and its binding induces large conformational changes. Activation by phenylalanine and phosphorylation act synergistically; the phosphorylated protein requires a lower concentration of phenylalanine for activation, and the phenylalanine-activated protein is more rapidly phosphorylated by cAMP-dependent protein kinase.

To understand the complex mechanisms regulating this important enzyme, we determined the crystal structure of rat PAH lacking the C-terminal, 24-residue tetramerization domain ($PAH_{1–428}$); the truncation of the C-terminal residues was necessary for reproducible crystallization, but it did not affect the regulatory and catalytic properties of the enzyme [13,14] (Figure 4). The structure revealed the expected two domains; a C-terminal catalytic domain, and an N-terminal regulatory domain. The most interesting obser-

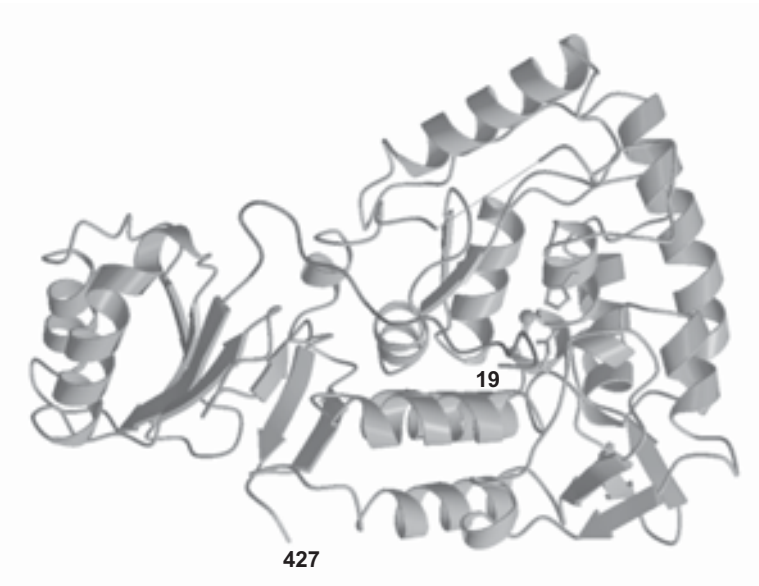

Figure 4. Structure of phenylalanine hydroxylase.[14] Ribbon diagram of a monomer of PAH_{1-428} is shown. The disordered residues 137–142 are indicated by a straight thin line. The side chains of the iron-binding residues (His285, His290 and Glu330) and the ferric ion are also shown.

vation was that the very N-terminal sequence comprising amino acids 19–29 reached into the active site of the catalytic domain. As was the case above with Impα, the sequence appeared to autoinhibit the enzyme.[6] We tested this autoinhibitory role of the N-terminal sequence by expressing a protein lacking the 29 N-terminal amino acids ($PAH_{30–428}$) and confirmed that $PAH_{30–428}$ was constitutively active (*i.e.* it does not require phenylalanine activation) and showed an altered structural response to phenylalanine.[15] Similar results were obtained using PAH lacking the first 26 residues.[16]

Another surprising observation revealed by the structure of $PAH_{1–428}$ was that residues 1–18, containing the phosphorylation site Ser16, showed no defined structure in both phosphorylated and un-phosphorylated forms;[14] this was difficult to reconcile with the established role of phosphorylation in activating the enzyme.[11] We used nuclear magnetic resonance (NMR) to follow the dynamics of the N-terminal mobile region. Our results confirm that this region is mobile in absence of phenylalanine, but a significant loss of mobility is observed for a portion of the sequence after the addition of phenylalanine, suggesting that upon activation the N-terminal sequence becomes associated with the folded core of the molecule (unpublished results). Our results suggest a model where the binding of phenylalanine to its regulatory site causes conformational changes, during which the N-terminal sequence moves away from the active site, with phosphorylation aiding this transition through stabilizing the phenylalanine-activated form. However, a structural characterization of the various ligand-bound states will be required for a complete understanding of the regulation of PAH.

MECHANISM OF RETROVIRAL MEMBRANE FUSION

Enveloped viruses, including the human pathogenic viruses such as influenza virus and the retroviruses HIV (human immunodeficiency virus) and HTLV (human T cell leukemia virus), fuse their membranes with host cell membranes to enable the transfer of their genome into the host cell.[17] Envelope glycoproteins (Env) are responsible for the attachment and subsequent membrane fusion of virions and infected cells with target cells. Env comprises a receptor-binding subunit (SU) associated with a transmembrane protein (TM). Envelope glycoproteins are synthesized as precursors and are processed in the Golgi apparatus to yield a mature functional SU/TM complex. The mature Env proteins are incorporated into budding virions at the plasma membrane. Retroviral entry into cells follows SU-mediated attachment to cellular receptors and TM-mediated fusion between the viral envelope and target cellular membrane. Cell surface-localized Env proteins

can also mediate cell-to-cell retrovirus transmission by fusion between infected cells and target cells.

We used a combination of crystal structure analysis and site-directed mutagenesis to study the Env protein from HTLV type 1 (HTLV-1). HTLV-1 is associated with various diseases including adult T-cell leukemia/lymphoma, HTLV-1-associated myelopathy/tropical spastic paraparesis, uveitis and infectious dermatitis of children.[18] We determined the crystal structure of the HTLV-1 TM ectodomain,[19] and used *in vitro* mutagenesis to test the functional role of structural elements observed in the gp21 structure.[20]

The ectodomain of gp21, the TM from HTLV-1, was crystallized as a maltose-binding protein (MBP) chimera and the MBP domain was used to solve the structure by the method of molecular replacement. The structure of gp21 comprises an N-terminal trimeric coiled coil, an adjacent disulfide-bonded loop that stabilizes a chain reversal, and a C-terminal sequence that packs against the coil in an extended anti-parallel fashion (Figure 5). The structure reveals both similarities and differences with TM fragments of other retroviruses such as human immunodeficiency virus (HIV), orthomyxoviruses such as influenza virus, a filovirus Ebola virus and a paramyxovirus, simian virus 5.[17] All these structures are long rods containing a central trimeric coiled coil and a chain reversal, which would imply that the hydrophobic N-terminal fusion peptide responsible for host cell membrane attachment, and the C-terminal transmembrane sequence anchored in the viral envelope, are juxtaposed at one end of the rod. These observations sug-

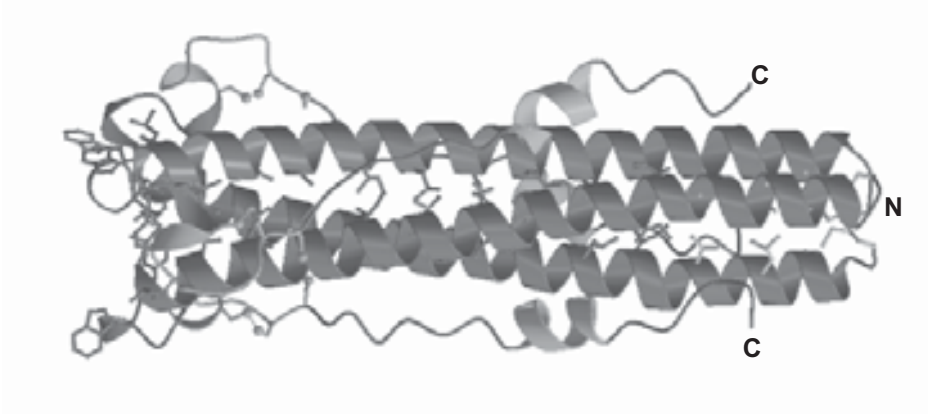

Figure 5. Structure of HTLV-1 gp21.[19] Ribbon diagram of the HTLV-1 gp21 ectodomain trimer is shown. The side chains of selected residues are also shown. The N- and the C-termini are indicated; the fusion peptides would be located at the N-termini, and the membrane anchors would be located at the C-termini. The maltose binding protein fusion partner used for crystallization is not shown.

gest that these structures represent a fusion-activated or post-fusion conformation. Our structure, jointly with all other available information, suggests a model for the mechanism of membrane fusion.[19] In this model, the SU/TM complex initially exists in a metastable, pre-fusion conformation, where the N-terminal fusion peptide and the C-terminal membrane anchor are at different ends of the complex. Host cell receptor binding by the SU subunit (*e.g.* in retroviruses) or a pH change (*e.g.* in influenza virus) triggers conformational changes, during which the N-terminal fusion peptide binds to the host cell membrane, followed by an anti-parallel packing of the C-terminal segment against the exterior of the coiled coil to provide energy for the relocation and juxtaposition of the C-terminal transmembrane anchors with the N-terminal fusion peptides. This rearrangement causes a perturbation of the two membranes that results in their subsequent fusion.

We further studied the structural basis of membrane fusion by assessing the functional consequences of mutating residues within the key structural elements of gp21 that are conserved in retroviral and filoviral TM hairpins.[20] We identified functional roles for a hydrophobic structure that caps the N-terminus of the central coiled coil, the region of chain reversal at the C-terminus of the coiled coil, and the sites of interaction between the C-terminal ectodomain segment and the coiled coil. We find that membrane fusion activity requires the stabilization of the N- and C-termini of the central coiled coil by a hydrophobic N-cap and a small hydrophobic core, respectively. A conserved Gly-Gly hinge motif preceding the disulfide-bonded loop, a salt bridge that stabilizes the chain reversal region and interactions between the C-terminal segment and the coiled coil are also critical for fusion activity. Our data support a model whereby the chain reversal region transmits a conformational signal from receptor-bound SU to induce the fusion-activated helical hairpin conformation of TM.

The Env-mediated process of membrane fusion has obvious parallels with the process of intrasteric regulation. While in the systems under intrasteric control an autoregulatory sequence autoinhibits the function of the protein through binding to its active site, in viral TM proteins the C-terminal sequence analogous to an autoregulatory sequence binds to the coiled coil to drive membrane fusion.

CONCLUSIONS

Protein-protein interactions and protein regulation are essential components of the complex networks functioning in cells, and understanding the molecular basis of these processes is crucial for piecing together a detailed map of cellular pathways. The work by our group has focused on a few spe-

cific systems where we can learn about the fundamental aspects of these processes, and we are delighted to see that our work has some general implications for a variety of other processes in the cell.

Acknowledgements. – I thank all my coworkers and collaborators who contributed to the studies described above, especially Bruno Catimel, Rob Center, Marcos Fontes, Ian Jennings, Bruce Kemp, Andy Poumbourios and Trazel Teh. This work was supported by the Wellcome Senior Research Fellowship in Medical Science in Australia, the Australian Research Council and the National Health and Medical Research Council.

REFERENCES

1. S. K. Burley, *Nature Struct. Biol.* **7** Suppl. (2000) 932–4.
2. C. Dingwall and R. A. Laskey, *Trends Biochem. Sci.* **16** (1991) 478–481.
3. S. R. Wente, *Science* **288** (2000) 1374–1377.
4. B. Kobe, *Nature Struct. Biol.* **6** (1999) 388–397.
5. E. Conti, M. Uy, L. Leighton, G. Blobel, and J. Kuriyan, *Cell* **94** (1998) 193–204.
6. B. Kobe and B. E. Kemp, *Nature* **402** (1999) 373–376.
7. B. Catimel, T. Teh, M. R. Fontes, I. G. Jennings, D. A. Jans, G. J. Howlett, E. C. Nice, and B. Kobe, *J. Biol. Chem.* **276** (2001) 34189–34198.
8. M. R. M. Fontes, T. Teh, and B. Kobe, *J. Mol. Biol.* **297** (2000) 1183–1194.
9. E. Conti and J. Kuriyan, *Structure* **8** (2000) 329–338.
10. D. A. Jans and S. Hübner, *Physiol. Rev.* **76** (1996) 651–685.
11. P. F. Fitzpatrick, *Annu. Rev. Biochem.* **68** (1999) 355–381.
12. C. R. Scriver, P. J. Waters, C. Sarkissian, S. Ryan, L. Prevost, D. Cote, J. Novak, S. Teebi, and P. M. Nowacki, *Hum. Mutat.* **15** (2000) 99–104.
13. B. Kobe, I. G. Jennings, C. M. House, S. C. Feil, B. J. Michell, T. Tiganis, M. W. Parker, R. G. H. Cotton, and B. E. Kemp, *Protein Sci.* **6** (1997) 1352–1357.
14. B. Kobe, I. G. Jennings, C. M. House, B. J. Michell, K. E. Goodwill, B. D. Santarsiero, R. C. Stevens, R. G. H. Cotton, and B. E. Kemp, *Nature Struct. Biol.* **6** (1999) 442–448.
15. I. G. Jennings, T. Teh, and B. Kobe, *FEBS Lett.* **488** (2001) 196–200.
16. G. A. Wang, P. Gu, and S. Kaufman, *Proc. Natl. Acad. Sci. USA* **98** (2001) 1537–1542.
17. P. Poumbourios, R. J. Center, K. A. Wilson, B. E. Kemp, and B. Kobe, *IUBMB Life* **48** (1999) 151–156.
18. A. Manns, M. Hisada, and L. La Grenade, *Lancet* **353** (1999) 1951–8.
19. B. Kobe, R. J. Center, B. E. Kemp, and P. Poumbourios, *Proc. Natl. Acad. Sci. USA* **96** (1999) 4319–4324.
20. A. L. Maerz, R. J. Center, B. E. Kemp, B. Kobe, and P. Poumbourios, *J. Virol.* 74 (2000) 6614–6621.
21. P. Kraulis, *J. Appl. Cryst.* **24** (1991) 946–950.
22. E. A. Merritt and M. E. P. Murphy, *Acta Crystallogr., Sect. D* **50** (1994) 869–873.

SAŽETAK

Intramolekulske samoregulirajuće sekvencije u proteinima

Bostjan Kobe

Općesvjetska inicijativa strukturne genomike obećava priskrbiti nam trodimenzijske strukture većine značajnih proteina za idućih deset godina. Ipak, ta inicijativa u početnim stadijima nije posebno usmjerena na strukturnu bazu interakcija protein-protein i proteinsku regulaciju, a to je oboje bitno za razumijevanje stanične funkcije proteina. Ovaj prikaz daje pregled naših istraživanja strukturnih temelja interakcija protein-protein i regulacije u nekoliko biološki važnih staničnih procesa. U žarištu je prepoznavanje jezgrovnih (nuklearnih) lokalizacijskih sekvencija (NLSs) s jezgrovnim importnim faktorom importinom-α i njegova regulacija. Također se diskutira o regulacijskoj interakciji proteina retrovirusne ovojnice i enzima fenilalaninhidroksilaze (PAH). Svim je tim sustavima zajednički mehanizam intrasteričke regulacije koji je određen aktivnim mjestom.