Diss. ETH No. 11579

Structure of Serum Response Factor core bound to DNA

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZÜRICH

for the degree of
Doctor of Natural Science

presented by

LUCA PELLEGRINI

Laurea in Chimica
born July 8th, 1966
Italian citizen

Prof. T. J. Richmond, Examiner
Prof. K. Wüthrich, Co-examiner

Zürich, 1996

Summary:

The human protein Serum Response Factor (SRF) regulates transcription of the "immediate-early" gene family and of muscle cell-specific genes like the actins. It binds as a dimer to DNA promoter sequences that contain the consensus decamer $CC(A/T)_6GG$, named CArG box. A core domain of about 90 amino acids is necessary and sufficient for the activities of DNA binding, dimerization and recruitment of the accessory factors that cooperate with SRF to activate transcription. Within the core domain is there a 58 amino acids DNA binding motif, called the MADS box, that has been identified in a large number of regulatory proteins from the animal and the plant kingdoms.

This work describes the subcloning of SRF's core domain into the pET3a expression vector, its overexpression in *E. coli* and purification to homogeneity, its cocrystallization with a DNA sequence containing a CArG box, and determination of the X-ray structure of the SRF core-DNA complex using the multiple isomorphous replacement method.

The structure of the SRF core-DNA complex shows that the members of the MADS family of transcription factors recognize their DNA site via a structurally novel DNA binding motif. In the complex, the two fold axis of the protein dimer coincides with the dyad axis of the DNA . The principal DNA binding element is an antiparallel coiled coil of two α-helices that sits on top of the minor groove and is aligned roughly parallel to it. The DNA molecule wraps around the basic N-termini of the helices, allowing contacts with the phosphates and the edges of the bases to take place. The N-terminal tails of the SRF core depart from the coiled coil and penetrate in the A/T rich minor groove, making important interactions with the DNA there. The C-terminal part of the MADS box is an hydrophobic β-hairpin that, in the protein dimer, pairs with the same element from the other monomer, forming a four-stranded β-sheet that covers the side of the coiled coil opposite the DNA. The C-terminal ends of the SRF core fold into

irregular coils followed by two short α-helices that close off the upper surface of the β-sheet, excluding it from any contact with the solvent. Overall, the SRF core has a compact, slab-like appearance. The DNA double helix is severely bent towards the major groove in direction of the protein at both ends of the CArG box, while the central, A/T rich part of the CArG box shows a very narrow minor groove. The SRF core exploits these conformational features of its recognition site to achieve sequence-specific binding, with only one base-specific contact in the major groove within the CArG box.

The three dimensional structure of the SRF core-DNA complex expands our knowledge of the general principles governing protein-DNA interaction and provides the necessary structural basis to understand how different members of the MADS box family can recognize DNA sequences that are similar but different.

Zusammenfassung:

Das menschliche Protein "Serum Response Factor" (SRF) reguliert die Transkription der "immediate-early" Genfamilie und muskelzell-spezifischer Gene. Es bindet als Dimer an DNA Promotor Sequenzen mit dem Konsensus-Dekamer $CC(A/T)_6GG$, CArG Box genannt. Eine Kern-Domäne von ungefähr 90 Aminosäuren ist notwendig und genügend für DNA-Bindung, Dimerisierung und Rekrutierung zusätzlicher Faktoren, die zusammen mit SRF die Transkription aktivieren. Diese Kern-Domäne enthält ein 58 Aminosäuren langes DNA bindendes Motiv, MADS Box genannt, das in einer grossen Zahl von regulatorischen Proteinen aus dem Tier- und Pflanzenreich identifiziert wurde.

Die vorliegende Arbeit beschreibt die Subklonierung der SRF Kern-Domäne in den Expressions-Vektor pET3a, die Überexpression in *E.coli* und die Aufreinigung, das Cokristallisieren mit einer CArG Box enthaltenden DNA Sequenz und die Bestimmung der Röntgenkristallstruktur des SRF Kern-Domäne/DNA Komplexes mittels MIR.

Die Struktur der SRF Kern/DNA Komplexes zeigt, dass die Mitglieder der MADS Transkriptionsfaktoren-Familie ihre DNA Bindestellen über ein strukturell neues DNA bindendes Motiv erkennen. Im Komplex fällt die Symmetrieachse (C2) des Proteins mit derjenigen der DNA zusammen. Das DNA bindende Element besteht aus einem antiparallelen "coiled coil" zweier alpha-Helices, die über der "minor groove" sitzen, ungefähr parallel zu ihr. Die DNA wickelt sich um die N-Termini der Helices, was Kontakte mit den Phosphatgruppen und den Rändern der Basen erlaubt. Die N-terminalen Stücke des SRF Kerns gehen vom "coiled coil" aus und dringen in die A/T reiche "minor groove" ein, wobei sie dort wichtige Kontakte zur DNA machen. Der C-terminale Teil der MADS Box ist eine hydrophobe beta-Haarnadel, die im Dimer mit der entsprechenden Haarnadel des anderen Monomers ein 4-strängiges beta-Blatt bildet. Dieses deckt die der DNA gegenüberliegende Seite des "coiled coil" ab. Die C-

terminalen Enden des SRF-Kerns falten sich zu unregelmässigen "Coils", gefolgt von zwei kurzen alpha-Helices. Diese schliessen das obere Ende des beta-Blattes ab und verhindern so jeglichen Lösungsmittelkontakt. Im grossen und ganzen hat der SRF Kern eine kompakte scheiben-ähnliche Form. Die DNA Doppelhelix ist an beiden Enden der CArG Box stark zur "major groove" in Richtung des Proteins gebogen, während der zentrale A/T reiche Teil der CArG Box eine sehr enge "minor groove" besitzt. Der SRF Kern nutzt diese Konformationsbesonderheiten seiner Erkennungsstelle für sequenz-spezifisches Binden mit nur einem basen-spezifischen Kontakt in der "major groove" der CArG Box.

Die dreidimensionale Struktur des SRF Kern/DNA Komplexes erweitert unser Wissen über allgemeinen Prinzipien, welche Protein-DNA Wechselwirkungen lenken. Sie stellt die notwendigen strukturellen Grundlagen zur Verfügung um zu verstehen, wie verschiedene Mitglieder der MADS Box Familie ihre ähnlichen aber unterschiedlichen DNA Bindestellen erkennen können.

Riassunto:

La proteina umana Serum Response Factor (SRF) regola la trascrizione della famiglia di geni "immediate-early" e di geni del tessuto muscolare come le actine. Si lega sotto forma di dimero a sequenze di DNA promotore che contengono il decamero $CC(A/T)_6GG$, chiamato CArG box. Un dominio di circa 90 amminoacidi (core domain) e' necessario e sufficiente per le attivita' di complessamente del DNA, dimerizzazione e reclutamento dei fattori ternari che cooperano con SRF per attivare la trascrizione. Dentro il "core domain" e' contenuta una sequenza di 58 amminoacidi capace di legare il DNA, chiamata MADS box, che e' stata identificata in un grande numero di fattori di trascrizione del regno animale e vegetale.

Questo lavoro descrive il subclonaggio del core domain del SRF nel vettore di escpressione pET3a, la sua overexpressione in *E. coli* e purificazione fino ad omogeneita', la sua cocristallizzazione con una sequenza di DNA contenente un CArG box, e la determinazione della struttura a raggi x del complesso SRF core-DNA, per mezzo del metodo della sostituzione multipla isomorfa.

La struttura del complesso SRF core-DNA mostra che i membri della famiglia MADS di fattori di trascrizione si legano al loro sito nel DNA tramite un nuovo motivo strutturale. Nel compesso, l'asse binario del dimero proteico coincide con l'asse binario del CArG box del DNA. Il principale elemento responsabile per il complessamento del DNA e' un "coiled coil" di due alfa eliche situate sopra il solco minore del DNA e allineate in modo approssimativamente parallelo ad esso. La molecola di DNA si avvolge attorno ai basici segmenti N-terminali delle alfa eliche, permettendo contatti con i fosfati e le basi di avere luogo. Le code N-terminali del SRF core si dipartono dal "coiled coil" e penetrano nel solco minore del DNA, ivi facendo importanti interazioni. I segmenti C-terminali del MADS box formano una forcina beta idrofobica che, nel dimero, si allinea con lo stesso elemento dell'altro monomero, formando un foglio beta antiparallelo con quattro "strands", che copre il lato del "coiled coil" opposto al DNA.

Le estremita' C-terminali del SRF core si avvolgono in modo irregolare per poi terminare in brevi alfa eliche che sigillano la superficie superiore del foglio beta, escludendolo da ogni contatto con il solvente. La doppia elica del DNA e' severamente piegata verso il solco maggiore in direzione della proteina ad entrambe le estremita' del CArG box, mentre il tratto centrale, ricco in A/T, presenta un solco minore molto stretto. Il SRF sfrutta le caratteristiche conformazionali proprie di questa sequenza di DNA per legarla in modo specifico, mentre le interazioni specifiche per la sequenza di basi sono limitate ad una, all'interno del CArG box.

La struttura tridimensionale del complesso SRF core-DNA espande la nostra conoscenza dei principi che governano le interazioni fra proteine e DNA, e fornisce la base strutturale necessaria per capire come membri differenti della famiglia MADS di proteine si leghino a sequenze simili ma diverse.