Università degli Studi di Roma "La Sapienza"

Dottorato di Ricerca in Biochimica
XVII Ciclo (A.A. 2001-2004)

EVOLUZIONE MOLECOLARE E VERSATILITÀ STRUTTURALE DEGLI ENZIMI DIPENDENTI DAL PIRIDOSSALE-5'-FOSFATO CON RIPIEGAMENTO DI TIPO I

Dottorando
ALESSANDRO PAIARDINI

Docente guida
Prof. Francesco Bossa

Coordinatore
Prof. Paolo Sarti

DICEMBRE 2004

PRESENTAZIONE DEL LAVORO

1.1 Inquadramento della ricerca

Negli ultimi anni si è assistito ad una crescita esponenziale della quantità di informazione disponibile su macromolecole biologiche, sull'onda dell'avvio e della conclusione di diversi progetti genomici, il cui contributo ha aperto nuove prospettive nella comprensione dei processi evolutivi responsabili della diversità funzionale di molte famiglie proteiche (Sanchez *et al.*, 2002; Fig. 1.1). Una delle principali sfide a cui la comunità scientifica è chiamata a rispondere consiste quindi nel dare significato a questa enorme massa di informazione di sequenza, sfruttando contemporaneamente i progressi notevoli ottenuti nel campo della biologia strutturale (Fig. 1.2), al fine di comprendere la logica molecolare e la storia evolutiva del vivente.

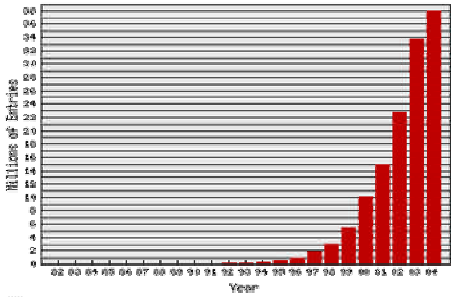


Fig. 1.1 Crescita della banca dati *GenBank* dal 1982 al 2004

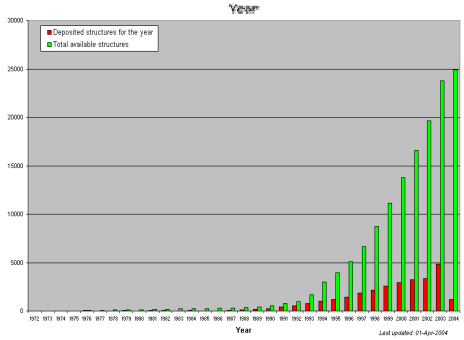


Fig. 1.2 Crescita della banca dati *Protein Data Bank* dal 1972 al 2004.

L'informazione biologica oggi disponibile costituisce la base di partenza grazie alla quale sarà forse possibile delineare con chiarezza il complesso rapporto che lega la sequenza amminoacidica alla struttura tridimensionale di una proteina. L'analisi di tale relazione, compiuta negli anni passati, ha posto spesso la ricerca di fronte a dei paradossi apparenti: nonostante gli esperimenti iniziati da Anfinsen (1973) e condotti nei decenni successivi dimostrino inequivocabilmente che la struttura primaria di una proteina contiene tutta l'informazione necessaria alla determinazione della sua conformazione nativa, tuttavia non sono affatto rari i casi di proteine omologhe che, a dispetto di una scarsa conservazione della sequenza amminoacidica, esibiscono strutture tridimensionali comparabili (Lesk & Chothia, 1980; Chothia & Lesk, 1986; Hill *et al.*, 2002); parimenti, proteine che condividono un'elevata identità di sequenza, presentano spesso caratteristiche chimico-fisiche e proprietà strutturali e catalitiche assai differenti (Britton *et al.*, 1985; Jin & Martin, 1999).

La presenza in natura di tali proteine solleva importanti questioni sul ruolo della struttura primaria nella determinazione del ripiegamento, della stabilità e delle proprietà proteiche: è possibile, ad esempio, estrapolare dal contesto della sequenza amminoacidica l'informazione minima necessaria al mantenimento della funzione e della conformazione nativa della struttura proteica, ignorando il rumore di fondo generato dalla minore pressione selettiva che agisce sulle regioni variabili? Considerato il ruolo essenziale svolto dai residui idrofobici nella determinazione del ripiegamento di una proteina, è ragionevole aspettarsi di riconoscere, in proteine che condividono uno stesso ripiegamento, un profilo idrofobico simile, la cui importanza è sottolineata da una maggiore conservazione, a livello evolutivo, dei residui che ne costituiscono il nucleo?

Le risposte a questi interrogativi permetteranno in futuro di ripercorrere l'evoluzione molecolare di famiglie e superfamiglie proteiche (Murzin *et al.*, 1995) e di capire come distinte proprietà catalitiche si siano evolute a partire da una comune struttura proteica ancestrale, attraverso successivi arrangiamenti genici, che comprendono plausibilmente mutazioni puntiformi, duplicazioni, fusioni, e riorganizzazione di domini funzionali. La piena comprensione dei meccanismi evolutivi sottesi al differenziamento ed all'adattamento molecolare all'ambiente avrà certamente un notevole impatto a livello applicativo, attraverso il disegno razionale di proteine con proprietà strutturali e funzioni non ancora esistenti in natura.

In questo ambito e nel tentativo di contribuire ad una risposta a tali interrogativi si colloca l'analisi *in silico* dell'evoluzione e della plasticità strutturale degli enzimi dipendenti dal piridossal-5'-fosfato (PLP), uno dei cofattori più versatili in natura (Fig. 1.3). Questi enzimi, che rivestono un ruolo fondamentale nel metabolismo degli aminoacidi, nella biosintesi di antibiotici, nella produzione di ammine attive ed in numerose altre vie biochimiche, sono capaci di catalizzare una vasta gamma di reazioni, che comprendono transaminazioni, racemizzazioni, α-decarbossilazioni, scissioni retroaldoliche, eliminazioni β e γ, e trasferimento di gruppi chimici (Jansonius, 1998; John, 1995).

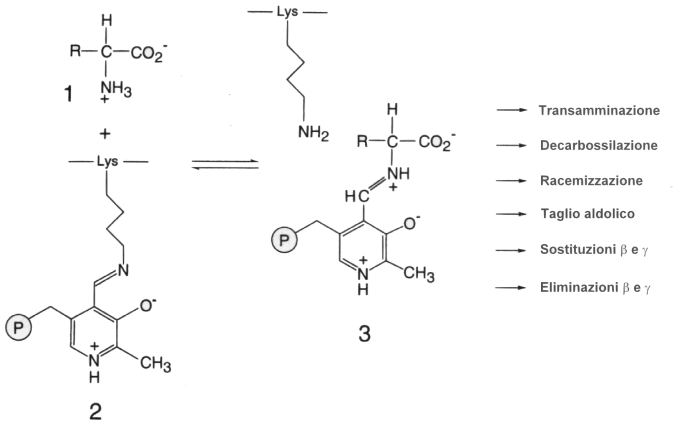


Fig. 1.3 Struttura chimica del PLP nelle due forme di aldimina interna (2) ed esterna (3), che si viene a formare a seguito del legame con il substrato (1) e tipi di reazioni catalizzate dagli enzimi PLP-dipendenti.

Sebbene siano state identificate sino ad oggi almeno cinque superfamiglie evolutivamente non correlate capaci di legare il PLP, ciascuna caratterizzata da un ripiegamento (*fold*) completamente differente, la superfamiglia funzionalmente più varia e che annovera il maggior numero di rappresentanti è conosciuta come superfamiglia dell'aspartato amminotrasferasi (Jansonius, 1998) o α (Mehta & Christen, 1998) o con ripiegamento di tipo I (Grishin *et al.*, 1995). Questa superfamiglia proteica è particolarmente adatta a rappresentare la plasticità strutturale ed evolutiva enzimatica: i membri di questo raggruppamento sono legati da una lunga storia di evoluzione divergente, a partire da un proto-enzima presente probabilmente nella cellula universale ancestrale, almeno 1.5 miliardi di anni fa (Mehta & Christen, 1998).

L'analisi compiuta in questo lavoro è stata indirizzata all'identificazione degli attributi strutturali comuni a tutti gli enzimi a PLP con ripiegamento di tipo I, plausibilmente responsabili della stabilità e del ripiegamento di questa superfamiglia. Questo studio ha permesso di riconoscere un insieme strutturale di contatti idrofobici conservati, condiviso da tutti i rappresentanti di questa superfamiglia. Questo assetto delinea le caratteristiche strutturali del proto-enzima capace di interagire con il PLP, ed evidenzia la presenza di un nucleo nel quale i residui che partecipano alle interazioni idrofobiche esibiscono una conservazione evolutiva preferenziale. Parallelamente, gli algoritmi e le metodologie di calcolo computazionale sviluppati per l'analisi *ad hoc* sono stati resi disponibili in rete e potranno essere utilizzati per l'analisi di ulteriori famiglie proteiche.

INTRODUZIONE

2.1 La superfamiglia di enzimi dipendenti dal PLP con ripiegamento di tipo I

Come accennato in precedenza (Paragrafo 1.1), tra le cinque superfamiglie di enzimi PLP-dipendenti, quella certamente più numerosa e meglio caratterizzata da un punto di vista biochimico, grazie soprattutto alle molteplici strutture cristallografiche risolte e depositate nelle banche dati, è rappresentata dagli enzimi con ripiegamento di tipo I (Grishin *et al.*, 1995). La struttura quaternaria minima fisiologicamente attiva negli enzimi appartenenti a questo ordine è l'omodimero, nonostante siano possibili complessi di ordine maggiore (Fig. 2.1).

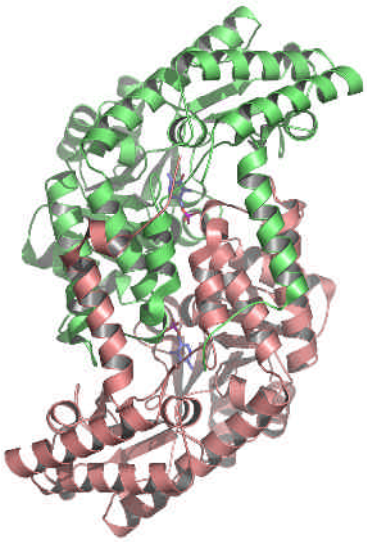


Fig. 2.1. Struttura del dimero della serina idrossimetiltransferasi di *E.coli*, codice PDB: 1DFO. Le due subunità sono colorate diversamente, in rosa e verde. Il PLP è in azzurro, con gli atomi di ossigeno e l'atomo di fosforo colorati in rosso e viola, rispettivamente.

Ciascuna subunità si ripiega in due domini: maggiore e minore. Il dominio maggiore contiene sette filamenti β nella sua porzione centrale a formare un foglietto β. Il minore, invece, si estende fino alla porzione C-terminale della catena e presenta quattro filamenti β circondati da α eliche su di un lato. La catena polipeptidica, in corrispondenza del dominio N-terminale, non ha un ripiegamento comune nella famiglia (Fig. 2.2).

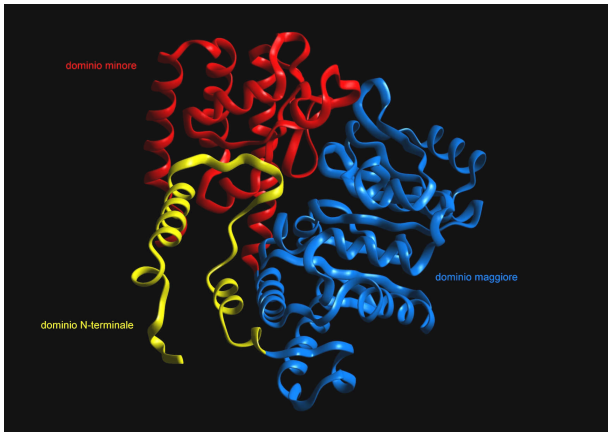


Fig. 2.2. Struttura del monomero a nastro della serina idrossimetiltrasferasi di *E. coli* colorata per domini (giallo: N-terminale; celeste: maggiore; rosso: minore).

Il PLP si trova posizionato all'interno di una tasca idrofobica, situata a sua volta all'interfaccia tra le due subunità del dimero. In assenza di substrato, il PLP si lega covalentemente ad un residuo di lisina che si trova nella porzione N-terminale di una α elica al dominio maggiore, a formare la così detta "aldimina interna"; in presenza di substrato, l'aldimina interna è convertita ad "aldimina esterna", attraverso la formazione di un intermedio a diammina geminale (Fig. 1.3). Il gruppo fosfato, invece, è stabilizzato con legami idrogeno dalla porzione N-terminale di un'altra α elica sul lato opposto rispetto ai foglietti β. La catena laterale di un residuo aromatico costringe l'anello aromatico del PLP ad impilarsi nei foglietti β. L'atomo di

azoto protonato N^1 dell'anello piridinico interagisce con un residuo di acido aspartico che ha il ruolo di stabilizzare la carica positiva presente nell'anello piridinico (Fig. 2.3), in accordo con l'ipotesi di Dunathan (1966), che spiega le proprietà catalitiche del PLP (Fig. 2.4).

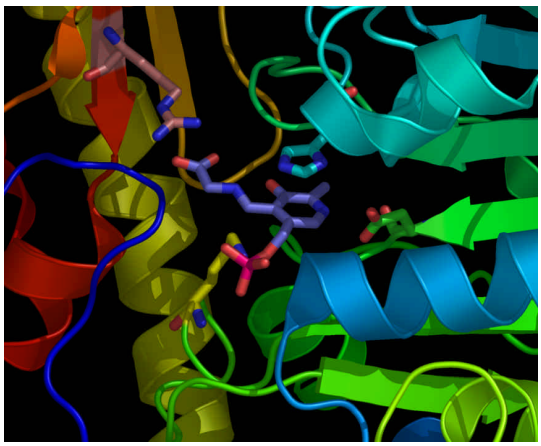


Fig. 2.3. Sito attivo della serina idrossimetiltrasferasi di *E. coli*. Il PLP legato nella forma di aldimina esterna con la glicina (PLG), è mostrato in celeste. Sono mostrate le catene laterali dei residui descritti nel testo (lisina, gialla; acido aspartico, verde; istidina, azzurro).

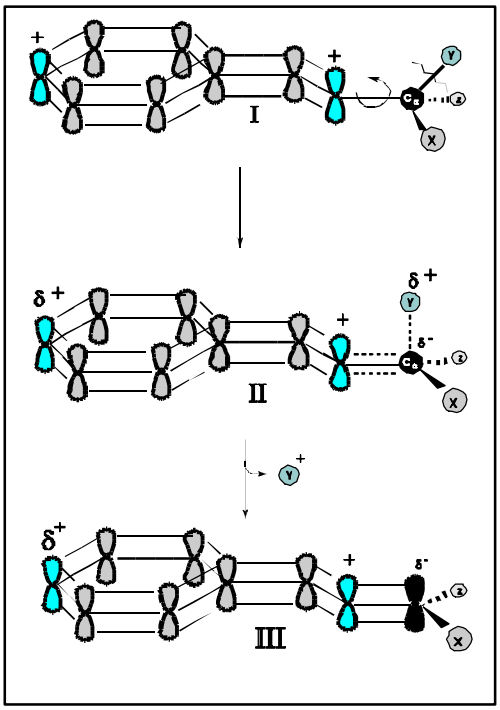


Fig. 2.4 Ipotesi di Dunathan per spiegare le proprietà chimico-fisiche del PLP. Il legame scisso viene a trovarsi perpendicolare al piano dell'anello del PLP. In questo modo la carica negativa che si forma sul carbonio α viene dispersa per risonanza con gli orbitali atomici non leganti del cofattore.

2.2 Le sottoclassi della famiglia con ripiegamento di tipo I

La superfamiglia degli enzimi PLP-dipendenti con ripiegamento di tipo I è stata ulteriormente suddivisa in famiglie. La ripartizione non risulta al momento univoca, ma degne di nota sono: la classificazione di Christen (Mehta & Christen, 1998), basata sulle caratteristiche funzionali e sulle informazioni di struttura primaria; quella di Käck (Schneider *et al.*, 2000), costruita sulle informazioni strutturali. Nel paragrafo seguente adotteremo questa ultima classificazione.

2.3 La classificazione in sei famiglie

Nella sezione 2.1 sono state indicate le caratteristiche strutturali comuni agli enzimi di tipo I; ciononostante, le numerose strutture cristallografiche di queste proteine risolte hanno permesso di rilevare importanti differenze.

Successivamente ad un allineamento strutturale di undici enzimi (Käck *et al.*, 1999) è stata proposta la seguente ripartizioni in sei classi:

- Amminotrasferasi che ulteriormente si dividono:
    1. Amminotrasferasi I;
    2. Amminotrasferasi II
    3. Fosfoserina amminotrasferasi;
- Tirosina fenol-liasi;
- Cistationina β liasi;
- Ornitina decarbossilasi;
- Serina idrossimetiltransferasi;
- 3-ammino-5-idrossibenzoico acido sintasi.

È da notare che esiste un'evidente correlazione tra la struttura della porzione N-terminale della catena polipeptidica e la suddivisione in sottoclassi determinata mediante l'allineamento strutturale di questa superfamiglia.

La struttura del dominio N-terminale differisce tra le sottoclassi, ma rimane identica all'interno di ognuna di esse (Fig 2.5).

Nella famiglia delle amminotrasferasi I, i primi residui della porzione N-terminale si trovano sulla superficie dell'altra subunità; i restanti, invece, attraversano la tasca del sito attivo prima di unirsi al dominio minore; qui, oltre a costituire l'ingresso del sito attivo, hanno il ruolo di stabilizzare i foglietti β presenti al C-terminale.

Le amminotrasferasi di classe II presentano un meandro β seguito da una α elica alla porzione N-terminale, importante nella stabilizzazione del dimero.

Il dominio N terminale nella fosfoserina amminotrasferasi è di soli quindici residui e forma un unico foglietto β che si unisce al dominio maggiore.

Nella sottoclasse della tirosina fenol-liasi questo dominio si estende con una α elica sulla regione cardine dell'enzima, e copre uno dei foglietti β all'estremità C-terminale. In tal modo, l'ingresso del sito attivo risulta più aperto. Altro ruolo di questo dominio è quello di mediare la formazione del tetrametro con un braccetto che si estende fuori dal sito attivo.

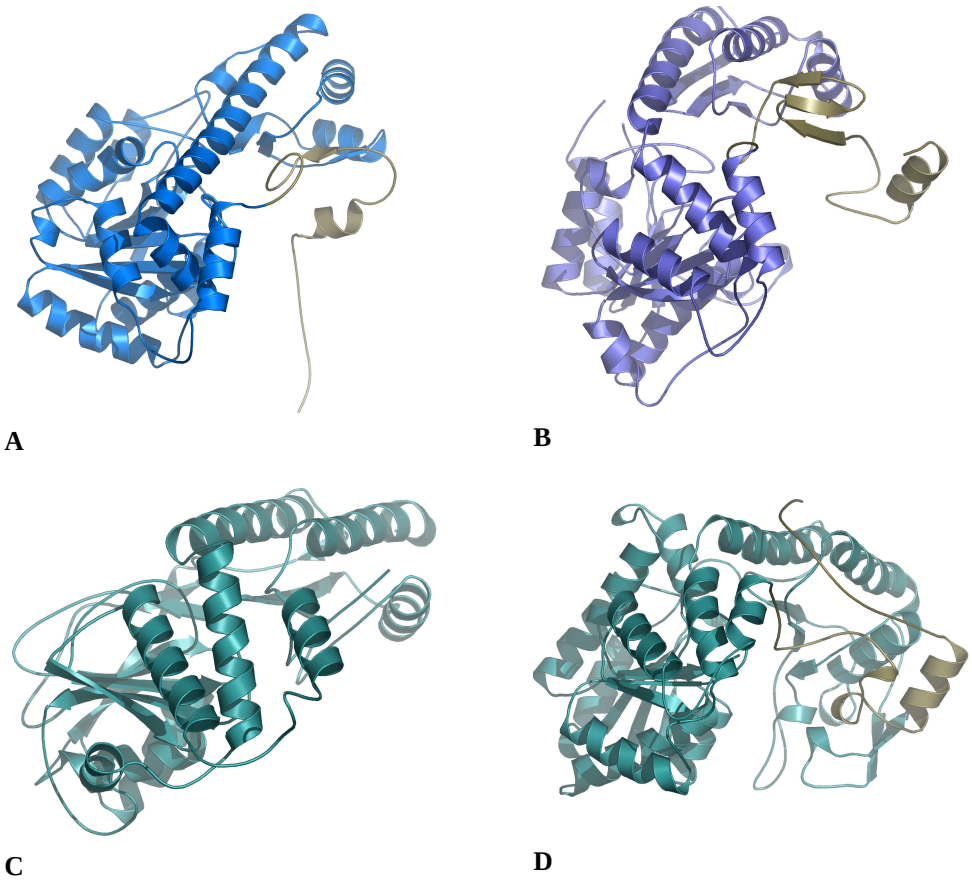
Nella cistationina β liasi il dominio N-terminale ha un braccetto che si estende nel dominio C-terminale dell'altra subunità del dimero ed interagisce con esso mediante alcune interazioni idrofobiche. Il dominio N-terminale, assieme con il dominio C-terminale, partecipa al legame del cofattore ed alla formazione dell'ingresso del sito attivo (Clausen *et al.*, 2000).

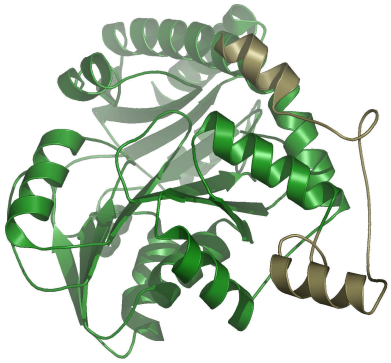
Il dominio N-terminale dell'ornitina decarbossilasi costituisce un dominio con una struttura tridimensionale a se stante, collegato al motivo legante il PLP da una lunga ansa.

La serina idrossimetiltrasferasi presenta una porzione N-terminale di circa cinquanta residui, che ha il ruolo di mediare i contatti tra le due subunità del dimero.

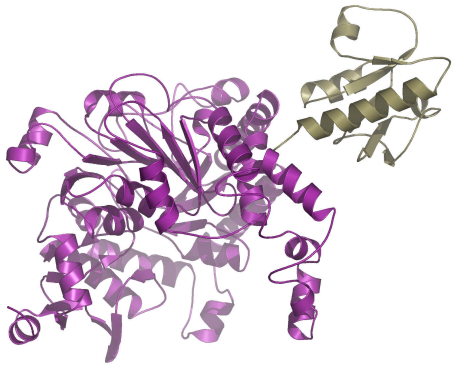
La classe della 3-amino-5-idrossibenzoico acido sintasi ha la porzione N-terminale che si collega direttamente al dominio maggiore.

Fig. 2.5 (pagine seguenti). Confronto tra le strutture dei monomeri delle sei sottoclassi degli enzimi PLP dipendenti con ripiegamento di tipo I, viste dallo stesso angolo di osservazione per enfatizzare le differenze di ripiegamento del dominio N-terminale colorato in oro. A) aspartato amminotrasferasi. PDB: 1ARS. B) acido diaminopelargonico sintasi. PDB: 1QJ5. C) fosfoserina aminotransferasi. PDB: 1QJ5. D) triptofano-indolo liasi. PDB: 1AX4. E) cistationina β liasi. PDB: 1CL1. F) ornitina decarbossilasi batterica. PDB: 1ORD. G) serina idrossimetiltransferasi. PDB: 1DFO. H) acido 3-amino-5-idrossibenzoico sintasi. PDB: 1B9H.

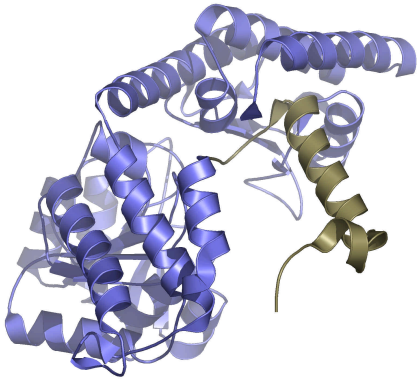




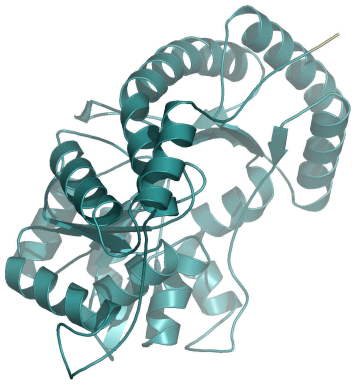
E



F



G



H

SCOPO DELLA RICERCA

Per la sua lunga storia evolutiva, la enorme variabilità strutturale e funzionale, e l'estesa caratterizzazione biochimica e strutturale, la superfamiglia degli enzimi PLP-dipendenti con ripiegamento di tipo I rappresenta un paradigma eccellente per capire come proteine omologhe, che condividono lo stesso ripiegamento, ma fortemente dissimili a livello di struttura primaria, si siano evolute a partire da un'impalcatura proteica comune, necessaria al legame del PLP. L'obiettivo principale di questo lavoro è stato quello di studiare le relazioni strutturali ed evolutive di questa superfamiglia, attraverso la sovrapposizione e l'allineamento di tutte le strutture cristallografiche di questi enzimi, depositate nelle banche dati, e la successiva identificazione delle regioni strutturalmente conservate (SCRs) e dei contatti idrofobici conservati (CHCs). A tal fine, sono stati sviluppati algoritmi e metodologie di calcolo computazionale, le cui prestazioni sono state preliminarmente valutate su famiglie proteiche estesamente studiate e caratterizzate in precedenza. Gli algoritmi sono stati poi implementati in programmi (CAMPO, SCR_FIND e CHC_FIND) messi a disposizione della comunità scientifica attraverso il sito *Web* http://schubert.bio.uniroma1.it.

Si è tentato, inoltre, di individuare un eventuale motivo topologico ancestrale legante il PLP, a partire dal quale si sono evoluti enzimi con specificità di reazione e substrato differenti, e di utilizzarlo come sonda per la ricerca nelle banche dati di altre famiglie enzimatiche contenenti questo motivo, con l'obiettivo di delineare eventuali caratteristiche strutturali comuni.

L'iniziale allineamento multiplo di strutture è stato successivamente esteso attraverso l'aggiunta di sequenze omologhe agli enzimi per i quali la

corrispondente struttura tridimensionale è già nota. A questo punto è stata condotta un'analisi evolutiva sull'allineamento finale di 921 sequenze, al fine di identificare i residui evolutivamente conservati (ECRs) e di stabilire la loro relazione con SCRs e CHCs. Infine, il ruolo giocato dai residui conservati nella stabilizzazione della struttura nativa ed il loro possibile coinvolgimento nel meccanismo di ripiegamento proteico sono stati discussi alla luce degli studi più recenti sugli enzimi PLP-dipendenti.

MATERIALI E METODI

4.1 Allineamenti strutturali

E' stata inizialmente condotta una ricerca di membri rappresentativi della superfamiglia di enzimi PLP-dipendenti di tipo I, dei quali è stata precedentemente risolta la relativa struttura tridimensionale, al fine di ottenere un insieme non ridondante sul quale condurre l'analisi. Utilizzando il sistema di classificazione di diverse banche dati (SCOP [Murzin *et al.*, 1995], CATH [Orengo *et al.*, 2003] e MMDB dell'NCBI [Chen *et al.*, 2003]), è stato raccolto un insieme esaustivo di strutture cristallografiche, dal quale sono stati selezionati 27 rappresentanti, sulla base di un criterio gerarchico di valutazione: inizialmente, sono stati scartati enzimi ingegnerizzati che presentavano mutazioni; successivamente, in presenza di enzimi ortologhi, è stata mantenuta la struttura con il più alto grado di risoluzione. Infine, in presenza di valori comparabili di risoluzione, è stato considerato anche il fattore *R*. Tutte le strutture sono state prelevate dalla banca dati *Protein Data Bank* (PDB; Berman *et al.*, 2000).

L'allineamento multiplo strutturale di partenza è stato ottenuto utilizzando l'algoritmo di "estensione combinatoriale" (*combinatorial extension*), implementato nel programma CE (Shindyalov & Bourne, 1998). Il risultato così ottenuto è stato successivamente raffinato in maniera manuale: l'allineamento di ciascuna coppia di enzimi è stato controllato e, se necessario, modificato affinché diverse proprietà strutturali delle due proteine, come le strutture secondarie, i residui implicati nel legame del cofattore e le regioni idrofobiche corrispondessero nell'allineamento in

maniera ottimale. In alcuni casi di ambiguità, ad esempio in presenza di ampie inserzioni o delezioni amminoacidiche, per le quali non è stato possibile definire l'esatto allineamento, si è preferito ricorrere, come criterio guida, ai punteggi assegnati dalla matrice di scambio BLOsum62 (Henikoff & Henikoff, 1992). Al termine di questa fase, al fine di ottenere un campione minimamente ridondante, le strutture che mostravano più del 30% di identità di sequenza con qualsiasi altra struttura presente nel campione sono state scartate. Ciò ha portato ad ottenere un insieme di 23 rappresentanti di enzimi PLP-dipendenti di tipo I a struttura nota, che mostravano tra loro una percentuale di identità massima del 27% ed una RMSD (*Root Mean Square Deviation*) massima dei carboni α, nei confronti a coppie, di 4.2 Å.

4.2 Identificazione delle regioni strutturalmente conservate (SCRs)

L'allineamento strutturale ottenuto (Paragrafo 4.1) è stato utilizzato per identificare l'impalcatura comune e le regioni strutturalmente conservate (SCRs) tra i membri di questa superfamiglia. Le SCRs sono state definite come regioni della catena polipeptidica caratterizzate da: i) una conformazione locale simile; ii) una RMSD dei carboni α in posizioni strutturalmente equivalenti minore od uguale a 3.0 Å (Hill *et al.*, 2002); iii) la mancanza di *indels* (inserzioni e delezioni) in tutte le strutture considerate, e iiii) una composizione minima di tre residui consecutivi. Al fine di identificare le SCRs, è stato sviluppato in linguaggio C e Perl ed implementato in un programma disponibile in rete (SCR_FIND, http://schubert.bio.uniroma1.it/SCR_FIND) un algoritmo capace di estrarre le regioni che soddisfano i criteri sopra menzionati dalle coordinate

tridimensionali delle strutture sovrapposte e dai relativi allineamenti multipli di sequenza.

Per ogni posizione equivalente *i* dell'allineamento multiplo strutturale, SCR_FIND calcola un punteggio basato sull'RMSD dal centro di massa dei carboni α strutturalmente equivalenti ed una penalità arbitraria *GP*, che viene aggiunta per ogni *gap* trovato (N_{gaps}), in base alla seguente relazione:

$$SC_i = \sqrt{\frac{\sum_{j=1}^{N} (x_{ji} - \bar{x}_i)^2 + (y_{ji} - \bar{y}_i)^2 + (z_{ji} - \bar{z}_i)^2}{N}} + GP \cdot N_{gaps} \quad (1)$$

dove x_{ji}, y_{ji} e z_{ji} rappresentano le coordinate cartesiane del $j_{\text{-mo}}$ carbonio α in posizione *i* dell'allineamento e $\bar{x}_i$, $\bar{y}_i$ e $\bar{z}_i$ sono le coordinate del centro di massa calcolato sugli *N* atomi trovati in posizione *i*. Una finestra *w* di dimensione 3 posizioni è fatta scorrere lungo l'allineamento ed utilizzata per identificare le posizioni con punteggio ≤ 3.0 (in questo caso, essendo stato assegnato alla penalità *GP* un valore molto elevato, il punteggio di 3.0 coincide in pratica con una RMSD ≤ 3.0 Å, e la totale assenza di *indels*). Ogni volta che una SCR candidata viene identificata, *w* viene aumentata iterativamente di una posizione finché il punteggio non supera il valore di 3.0, o finché non viene raggiunta la fine dell'allineamento.

4.3 Identificazione dei contatti idrofobici conservati (CHCs)

I risultati ottenuti con SCR_FIND possono essere utilizzati da CHC_FIND (http://schubert.bio.uniroma1.it/CHC_FIND) per identificare i contatti idrofobici conservati nelle SCRs identificate. Questo programma si avvale dell'algoritmo di Drabløs (1999), che calcola la superficie idrofobica di contatto tra coppie di atomi non polari, per derivare la superficie idrofobica di contatto dei residui presenti nelle SCRs di tutte le strutture analizzate. I CHCs sono successivamente classificati sulla base della loro posizione (CHCs all'interno della stessa SCR e CHCs tra due diverse SCRs), il numero di strutture all'interno delle quali il contatto idrofobico è conservato e la superficie media di contatto apolare dei residui strutturalmente equivalenti di ogni struttura. Se due posizioni dell'allineamento multiplo strutturale, *x* ed *y*, presentano residui in contatto in almeno due strutture, allora un CHC candidato viene identificato. I CHCs sono successivamente classificati sulla base della loro intensità s_{xy}, definita come:

(2) $$s_{xy} = \frac{\sum_{i=1}^{N} A_i(x, y)}{N}$$

dove A_i rappresenta l'area di contatto apolare della struttura $i_{\text{-}ma}$ tra i residui in posizione assoluta *x* ed *y* dell'allineamento strutturale, ed *N* è il numero di strutture sovrapposte.

4.4 Estensione dell'allineamento multiplo strutturale

L'allineamento multiplo strutturale precedentemente ottenuto (Paragrafo 4.1) è stato successivamente esteso attraverso l'aggiunta di sequenze omologhe alle proteine a struttura nota. La ricerca di sequenze è stata condotta sulla banca dati NRDB (Holm & Sander, 1998), utilizzando il programma BLAST (Altschul *et al.*, 1997) ed ognuna delle 23 strutture sovrapposte come sonda. Quando applicabili, sono stati utilizzati i seguenti criteri per filtrare le sequenze ottenute:

- significatività dell'allineamento ≤0.0001 (Altschul *et al.*, 1997);
- percentuale minima di identità con la sequenza sonda >30%;
- percentuale massima di identità con qualsiasi altra sequenza dell'allineamento <80%;
- percentuale minima di posizioni allineate sulla sequenza totale >80%.

Le sequenze filtrate sono state allineate ognuna alla struttura corrispondente utilizzando il programma CLUSTALW (Thompson *et al.*, 1994). I 23 allineamenti multipli così ottenuti sono stati successivamente uniti, utilizzando come guida l'allineamento multiplo strutturale dei 23 enzimi di partenza (Pascarella & Argos, 1992). L'allineamento finale, comprensivo di 973 sequenze, è stato nuovamente controllato per eliminare ogni ridondanza. Al termine di questo ultimo passaggio, sono state ottenute un totale di 921 sequenze.

4.5 Identificazione dei residui evolutivamente conservati (ECRs)

L'identificazione dei residui evolutivamente conservati è stata condotta grazie allo sviluppo di un apposito algoritmo, implementato nel programma CAMPO (http://schubert.bio.uniroma1.it/CAMPO). Per misurare la conservazione evolutiva di ogni posizione dell'allineamento multiplo,

CAMPO assegna un punteggio formalmente simile a quello proposto da Karlin & Brocchieri (1996):

$$(3)\quad O_k = \frac{1}{\left(n(n-1)/2\right)}\left[\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\frac{Bscore_{kij}}{1\Big/2\left[\left(\sqrt{Bscore^2{}_{kii}}\right)+\left(\sqrt{Bscore^2{}_{kjj}}\right)\right]}\cdot\left(1-\frac{nid_{ij}}{nal_{ij}}\right)\right]}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(1-\frac{nid_{ij}}{nal_{ij}}\right)}\right]$$

dove O_k è il punteggio assegnato ad ogni posizione k dell'allineamento multiplo di sequenza, n è il numero di sequenze incluse nell'allineamento, i e j si riferiscono alla $i_{\text{-}ma}$ ed alla $j_{\text{-}ma}$ sequenza, rispettivamente, $Bscore_{kij}$, $Bscore_{kii}$ e $Bscore_{kjj}$ sono i punteggi assegnati agli scambi amminoacidici in posizione k tra la $i_{\text{-}ma}$ e la $j_{\text{-}ma}$ sequenza, in base ai valori riportati nelle matrici di scambio amminoacidiche BLOsum o PAM, nid_{ij} equivale al numero di residui identici e nal_{ij} è il numero di residui allineati tra la $i_{\text{-}ma}$ e la $j_{\text{-}ma}$ sequenza, rispettivamente.

Di conseguenza, per ogni possibile scambio in una particolare posizione dell'allineamento multiplo, viene calcolato un indice di conservazione normalizzato, basato sul punteggio di una matrice di mutazione. Dal momento che i punteggi della matrice per l'appaiamento di coppie identiche di amminoacidi variano a seconda dell'identità del residuo considerato, i punteggi di conservazione per posizioni invarianti dell'allineamento multiplo dipenderebbero dal tipo di residuo presente. Perciò, la normalizzazione si rende necessaria per evitare la presenza di punteggi differenti in posizioni

invarianti. A differenza del metodo proposto da Karlin & Brocchieri (1996), in questo algoritmo viene incorporato un sistema di pesatura nel quale ogni punteggio è corretto in base all'identità di sequenza tra le proteine che sono confrontate. In questo modo si assegna un peso maggiore alla conservazione di un residuo in sequenze lontane dal punto di vista evolutivo, rispetto alla stessa conservazione in sequenze vicine. Recentemente, nell'assegnazione del punteggio di conservazione evolutiva, sono stati adottati schemi di pesatura più sofisticati, basati su analisi filogenetiche. (Altschul *et al.*, 1997 e Armon *et al.*, 2001). Comunque, come discusso da Valdar (2002), gli schemi basati su alberi filogenetici richiedono spesso più assunzioni di quelli che sono basati direttamente sugli allineamenti di sequenza e possono introdurre ulteriore incertezza nel punteggio finale.

I valori della media O e la deviazione standard (SD) s per la distribuzione dei valori O_k vengono poi calcolati; la significatività R di ogni indice di conservazione dell'allineamento viene successivamente calcolata dividendo la differenza tra O_k e O per s.

Sebbene nell'analisi della conservazione evolutiva condotta sugli enzimi PLP-dipendenti di tipo I non ne sia stato fatto uso, i punteggi calcolati in ogni posizione dell'allineamento multiplo possono essere opzionalmente utilizzati come pesi per calcolare la conservazione evolutiva di una porzione di spazio di raggio arbitrario D, centrata su ogni atomo della struttura tridimensionale, attraverso l'applicazione di una tecnica ispirata alla teoria della percolazione (Harrison, 2001; Paragrafo 5.4):

(4)
$$R_{inew} = R_i + \frac{\sum_{j=1}^{n} \begin{cases} \frac{R_j}{d_{ij}} & \text{se } d_{ij} \leq \text{D} \\ 0 & \text{se } d_{ij} > \text{D} \end{cases}}{\sum_{j=1}^{n} \begin{cases} \frac{1}{d_{ij}} & \text{se } d_{ij} \leq \text{D} \\ 0 & \text{se } d_{ij} > \text{D} \end{cases}}$$

dove n è il numero di atomi della molecola, R_{inew} rappresenta il punteggio ricalcolato per l'atomo i, R_i e R_j rappresentano i punteggi iniziali assegnati agli atomi i e j, rispettivamente, corrispondenti al punteggio calcolato per i residui ai quali gli atomi appartengono, d_{ij} è la distanza tra i centri di massa dei residui ai quali gli atomi i e j appartengono, e D è un raggio arbitrario definito dall'utente, o limite di percolazione.

RISULTATI

5.1 Regioni strutturalmente conservate

Il campione collezionato, utilizzato in questo lavoro, comprende 23 strutture cristallografiche e 921 sequenze di enzimi PLP-dipendenti di tipo I, provenienti da organismi differenti, che nell'insieme comprendono i tre domini della vita, Eucarioti, Batteri ed Archaea (Tab. 5.I). Il livello di identità di sequenza tra le strutture sovrapposte ha garantito la completa copertura dei valori all'interno della così detta "zona crepuscolare" (Rost, 1999), con un intervallo compreso tra il 6% ed il 27% (media, 12%, deviazione standard [SD], ±3%). A dispetto della bassa identità di sequenza, questa superfamiglia enzimatica mostra una conservazione strutturale notevole, con una RMSD massima, misurata a coppie di strutture, di 4.2 Å (Fig. 5.1 A e B).

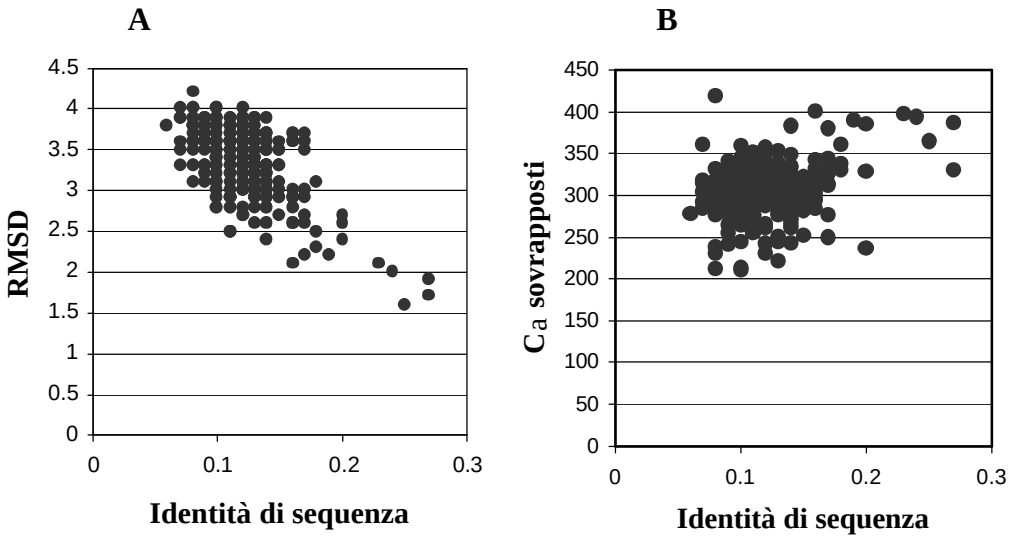


Fig. 5.1. (A) Distribuzione dell'RMSD e (B) numero di carboni α sovrapposti in funzione dell'identità di sequenza a coppie del campione analizzato. Ogni punto nel grafico rappresenta un confronto tra due strutture.

Tabella 5.I. Enzimi PLP-dipendenti utilizzati per l'analisi*

PDB	Descrizione dell'enzima	Organismo	Risoluzione	Valore *R*	Referenza	Min. % identità	N° di omologhi
1BJ4	serina idrossimetiltrasferasi	*H.sapiens*	2.65	0.210	Renwick et al., 1998	35	101
1BS0	8-amino-7-oxonanoato sintasi	*E.coli*	1.65	0.178	Alexeev et al., 1998	35	47
1FG3	istidinol fosfato amminotrasferasi	*E.coli*	2.20	0.198	Sivaraman et al., 2001	35	32
1JG8	treonina aldolasi	*T.maritima*	1.80	0.207	Kielkopf & Burley, 2002	40	21
1ECX	proteina nifs-like	*T.maritima*	2.70	0.207	Kaiser et al., 2000	40	101
1BJW	aspartato amminotrasferasi	*T.thermophilus*	1.80	0.215	Nakai et al., 1999	40	43
1D2F	proteina maly	*E.coli*	2.50	0.201	Clausen et al., 2000	40	2
1C7N	cistalisina	*T.denticola*	1.90	0.208	Krupka et al., 2000	30	35
1ELQ	L-cisteina/L-cistina C-S liasi	*Synechocystis*	1.80	0.198	Clausen et al., 2000	30	8
1DGD	dialchilglicina decarbossilasi	*B.cepacia*	2.80	0.178	Hohenester et al., 1994	35	98
1BJN	fosfoserina amminotrasferasi	*E.coli*	2.30	0.175	Hester et al., 1999	40	38
1AY4	aromatic amino acid amminotrasferasi	*P.denitrificans*	2.33	0.175	Okamoto et al., 1998	40	21
1DTY	adenosilmetionina amminotrasferasi	*E.coli*	2.14	0.196	Alexeev et al., 1998	40	49
2GSA	glutammato semialdeide amminomutasi	*C.biosynthesis*	2.40	0.183	Hennig et al., 1997	40	58
1B9H	3-ammino-5-idrossibenzoato sintasi	*A.mediterranei*	2.00	0.218	Eads et al., 1997	35	9
1AX4	triptofanasi	*P. vulgaris*	2.10	0.186	Isupov et al., 1998	40	15
1CL1	cistationina beta-liasi	*E.coli*	1.83	0.151	Clausen et al., 1996	40	23
1GTX	4-aminobutirrato amminotrasferasi	*Sus scrofa*	3.00	0.186	Storici et al., 1999	40	9
1JS6	dopa decarbossilasi	*Sus scrofa*	2.60	0.206	Burkhard et al., 2001	35	27
1ORD	ornitina decarbossilasi	*Lactobacillus sp.*	3.00	0.219	Momany et al., 1995	35	17
1QGN	cistationina gamma-sintasi	*N. tabacum*	2.90	0.201	Steegborn et al., 1999	35	108
1LK9	alliina liasi	*Allium sativum*	1.53	0.193	Kuettner et al., 2002	40	2
1MDX	arnb amminotrasferasi	*S.typhimurium*	1.96	0.206	Noland et al., 2002	35	57

* Sono stati utilizzati diversi criteri per ridurre la ridondanza nel campione utilizzato: sono state conservate solo sequenze significativamente simili (*E*-value ≤0.001); sono state rigettate sequenze con una percentuale di identità >80% rispetto alla sonda utilizzata, così come omologhi distanti (<30% di identità), per i quali l'incertezza nell'accuratezza dell'allineamento con le sequenze per le quali la struttura è nota è troppo alta. Inoltre, le sequenze con una lunghezza <80% rispetto alla sonda sono state rigettate in modo da ottenere nell'allineamento finale solo sequenze complete e non frammenti.

Nel tentativo di identificare le regioni comuni del nucleo proteico ed i residui coinvolti in ruoli strutturali e funzionali, lo studio è stato focalizzato sui segmenti di catena polipeptidica che conservano una conformazione simile della catena polipeptidica in tutte le strutture tridimensionali analizzate (SCRs, vedi Paragrafo 4.2), con l'esclusione delle regioni che invece differiscono marcatamente tra proteine differenti. Le SCRs sono state soggette a pressioni evolutive simili durante la diramazione di questi enzimi da un antenato comune; è quindi possibile che queste regioni contengano la maggior parte dei determinanti strutturali necessari al corretto ripiegamento proteico. Sono state identificate 17 regioni con una RMSD ≤ 3.0 Å, prive di inserzioni e delezioni (Fig. 5.2 e Fig. 5.3). I valori di RMSD per ogni posizione dell'allineamento sono compresi tra 0.8 Å in posizione 65 (la numerazione si riferisce alle posizioni mostrate in Fig. 5.2), e 3.6 Å in posizione 126 (Tabella 5.II). La figura 5.3 evidenzia la presenza di un'estesa organizzazione comune della catena polipeptidica intorno al PLP, responsabile del corretto posizionamento di residui precedentemente identificati come determinanti strutturali indispensabili al legame del cofattore (Grishin *et al.*, 1995). Cinque SCRs sono principalmente implicate nella costituzione di questa regione comune: una α-elica (α_3, che mostra una RMSD di 1.59 Å) e quattro filamenti β, che formano un foglietto β (β_6, β_9, β_{10} e β_{11}, con una RMSD di 1.76 Å, 1.54 Å, 1.41 Å e 1.52 Å, rispettivamente).

5.2 Residui evolutivamente conservati

Considerata la conservazione strutturale dei residui coinvolti nella formazione delle SCRs e la loro possibile rilevanza per la stabilità degli

enzimi PLP-dipendenti di tipo I, è stata condotta un'analisi per comprendere fino a che punto le loro proprietà chimico-fisiche, e di conseguenza il loro ruolo funzionale, è stato preservato durante l'evoluzione.

A tal fine è stata condotta una ricerca di sequenze omologhe a quelle delle proteine selezionate a struttura nota. Le sequenze ottenute sono state così utilizzate per estendere l'allineamento multiplo strutturale iniziale. Sono stati adottati diversi criteri per ridurre la presenza di ogni possibile ridondanza nel campione analizzato (Paragrafo 4.4): le sequenze con una percentuale di identità >80% con qualsiasi altra proteina già presente nel campione sono state scartate, così come le sequenze evolutivamente distanti (<30% di identità di sequenza), per le quali non era possibile assicurare un'elevata accuratezza dell'allineamento con le sequenze a struttura nota (Vogt *et al.*, 1995). Il numero di sequenze mantenute per ogni struttura è mostrato nella Tabella 5.I.

L'allineamento multiplo strutturale ottenuto dalla sovrapposizione delle strutture cristallografiche è stato poi utilizzato come guida per fondere i 23 allineamenti multipli di sequenza, per un totale di 921 sequenze non ridondanti. 376.573 dei 422.740 confronti a coppie di sequenze hanno riportato una percentuale di identità di sequenza nell'intervallo 0-20% (media 16%, SD ±6%); questo dato suggerisce che il campione utilizzato può rappresentare eventi evolutivi molto distanti (Fig. 5.4).

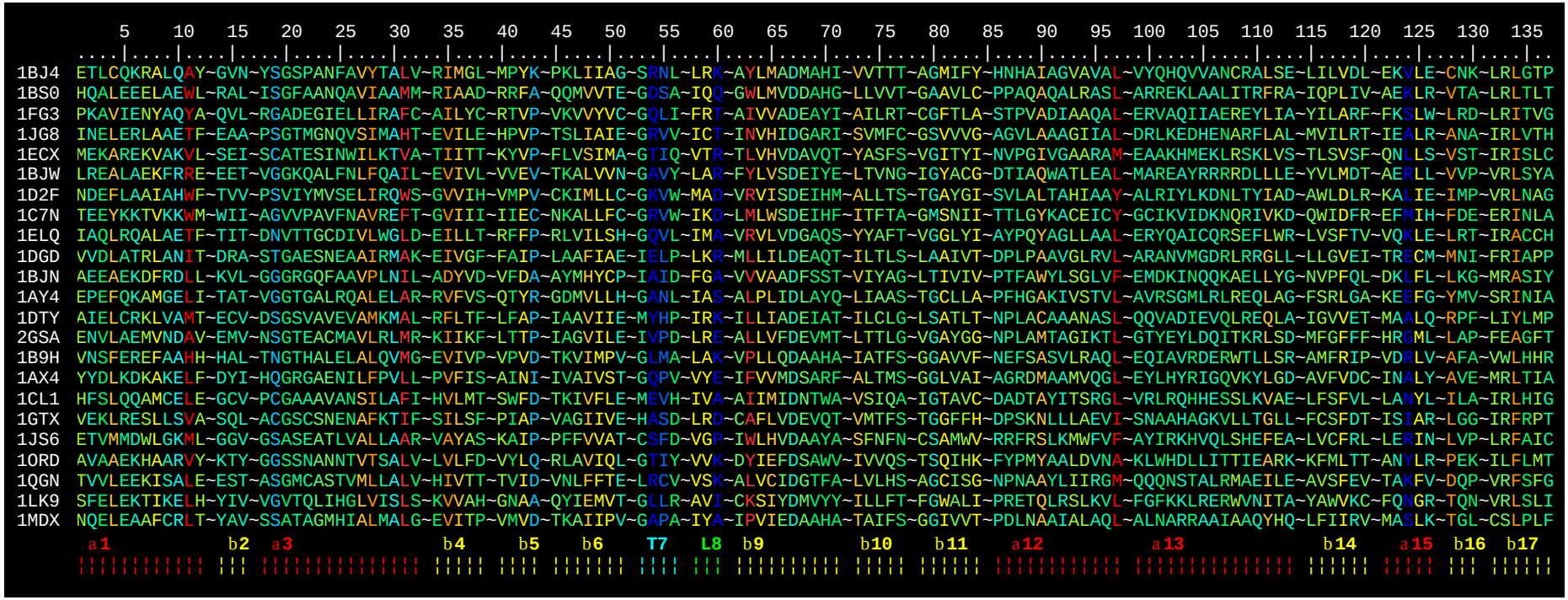


Fig. 5.2. Allineamento delle SCRs negli enzimi PLP-dipendenti di tipo I. Le SCRs sono rappresentate come blocchi separati da trattini. La linea in alto rappresenta la posizione assoluta dell'allineamento. Le colonne dell'allineamento sono colorate in accordo con lo schema di colori della Fig. 5.7 Ogni sequenza è identificata con il codice PDB della struttura corrispondente. I tratti nella porzione più bassa rappresentano elementi di struttura secondaria, indicati in base allo schema corrispondente: α, elica α; β, filamento β; L, ansa; T, Giro (*Turn*).

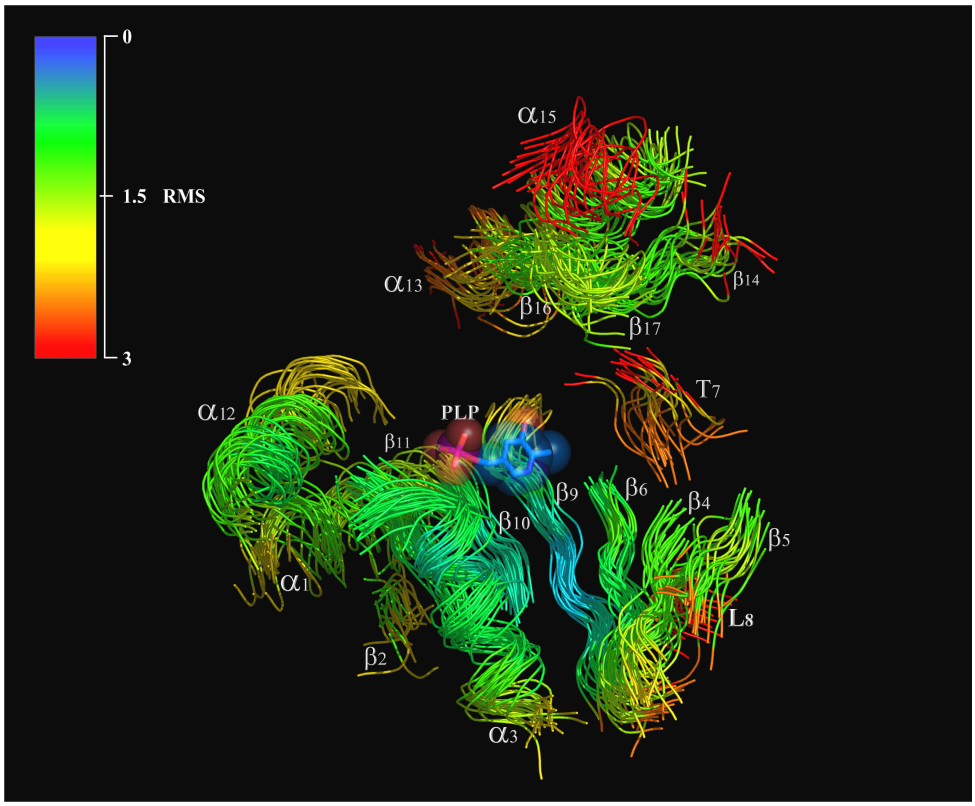


Fig. 5.3. Sovrapposizione delle SCRs trovate negli enzimi PLP-dipendenti con ripiegamento di tipo I. Le catene polipeptidiche delle 23 strutture sovrapposte sono mostrate come filamenti. Sono state individuate 17 regioni con una RMSD media ≤ 3.0 Å, prive di inserzioni e delezioni, e gli atomi corrispondenti colorati in accordo al valore di RMSD. Il PLP è mostrato in azzurro, con gli atomi di ossigeno in rosso, gli atomi di azoto in blu, ed il fosforo viola. Ogni SCR è indicata in base allo schema riportato in Fig. 5.2.

Dopo aver ottenuto l'allineamento multiplo di sequenza, è stato possibile applicare un metodo per l'identificazione dei residui evolutivamente conservati (ECRs). Dal momento che è stato dimostrato che in confronti di

allineamenti multipli di sequenza la matrice BLOsum62 è in grado di fornire risultati più accurati rispetto alle altre matrici (Vogt *et al.*, 1995), nello sviluppo di una metodologia per l'identificazione delle posizioni evolutivamente più conservate in un allineamento è sembrato opportuno adottare questa matrice, per l'assegnazione dei punteggi dei singoli scambi amminoacidici. Inoltre, è stato adottato anche uno schema di pesatura basato sull'identità delle sequenze analizzate, per incorporare nell'algoritmo una correzione della possibile ridondanza nelle posizioni delle sequenze simili e dei possibili errori dovuti alla distanza evolutiva delle sequenze (Paragrafo 4.5).

I risultati ottenuti per le SCRs, espressi in unità di deviazioni standard *R* dalla media di conservazione evolutiva, sono mostrati in Tabella 5.II. Il ruolo strutturale giocato dalle SCRs nel mantenimento del ripiegamento di questa superfamiglia enzimatica è riflesso dall'alta conservazione di sequenza delle posizioni corrispondenti dell'allineamento multiplo. I punteggi mostrati dalle SCRs si trovano infatti tutti al di sopra del valore medio di conservazione, con la sola eccezione del sito 14, che ha ottenuto un punteggio negativo. In particolare, i residui che interagiscono con il PLP sono i più conservati: l'aspartato 67, che interagisce con l'azoto dell'anello piridinico del PLP (Mehta & Christen, 1998), è stato trovato in 919 delle 921 sequenze allineate (le sole eccezioni sono l'8-amino-7-oxononanoato sintasi da *Mesorhizobium loti* e la Cistationina β-liasi da *Bifidobacterium longum*, codici *GenBank*: GI 13475018 e GI 23336039 [Holm & Sander, 1998] rispettivamente, nelle quali l'aspartato è sostituito con una glicina ed un'asparagina), ottenendo una significatività di 3.3 SD dal punteggio medio di conservazione.

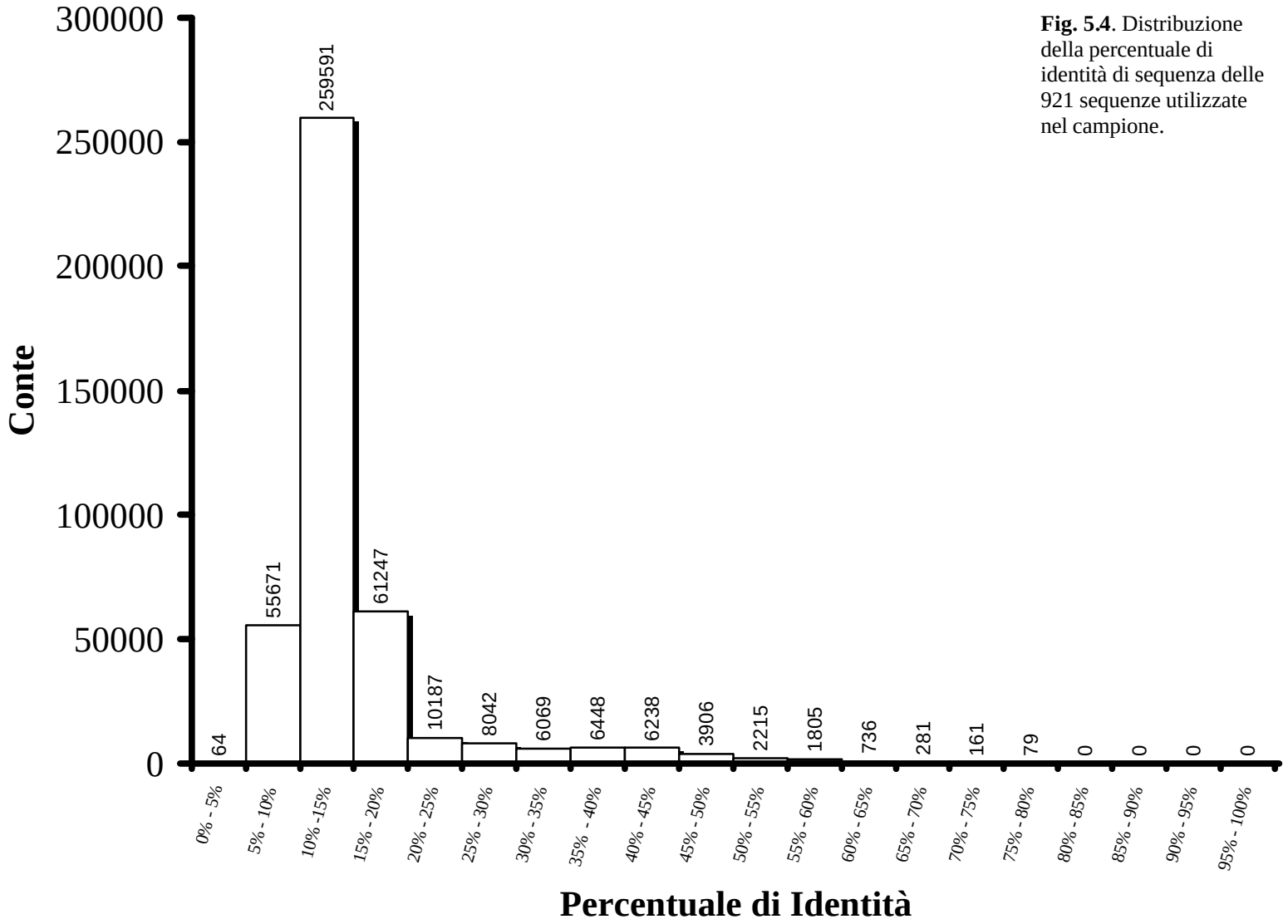


Fig. 5.4. Distribuzione della percentuale di identità di sequenza delle 921 sequenze utilizzate nel campione.

Un valore comparabile (R = 3.2) è stato osservato solo nel caso della lisina che forma la base di Schiff con il PLP, posta in una regione variabile tra le SCRs β_{10} e β_{11} (Christen & Mehta, 2001). Considerati insieme, questi due residui costituiscono la sequenza firma principale di questa superfamiglia di enzimi.

Altri siti coinvolti nelle interazioni con il cofattore od i substrati sono fortemente conservati: la posizione 70, ad esempio, che interagisce con l'ossigeno fenolico del PLP (R = 1.2), i residui che si impilano sulla faccia *re* e *si* del PLP (R = 1.6 ed R = 1.4, rispettivamente), la così detta "regione ricca in glicine" (posizioni 19, 20 e 21; R = 1.6, 1.9 ed 1.0, rispettivamente), il residuo legante il 5'-fosfato in posizione 77 (R = 1.5) e l'arginina impegnata in una interazione di tipo elettrostatico con il gruppo α carbossilico di molti substrati legati dagli enzimi di tipo I (sito 133, R = 2.0; Fig. 5.5).

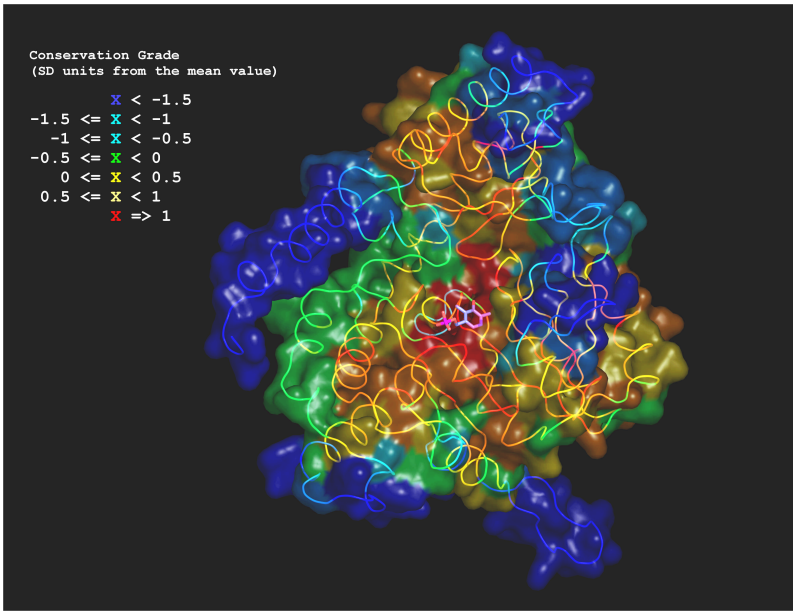


Fig. 5.5. Conservazione evolutiva indicata sulla catena polipeptidica e sulla superficie molecolare della serina idrossimetiltrasferasi da *H. sapiens*, utilizzata come rappresentante di questa superfamiglia enzimatica. I risultati ottenuti per le SCRs sono espressi in unità di SD dal valore medio di conservazione (R). Blu scuro corrisponde alla massima variabilità, rosso alla massima conservazione. Il PLP è mostrato in celeste. Gli atomi di ossigeno sono colorati in rosso, l'azoto in blu, ed il fosforo viola.

Tabella 5.II (Pagine seguenti) Attributi strutturali e di sequenza delle SCRs

SCR	RMS	SITO	RMS	Residui trovati in ogni posizione dell'allineamento multiplo strutturale*	Max. superficie di contatto apolare (Å²)	Punteggio di conservazione	Interazione con SCR, sito
		1	2.2	EHPIMLNTIVAEAEVYHVEATSN	15.9	0.49	12, 90
		2	2.2	TQKNERDEAVEPINNYFETVVFQ	6.1	0.46	1,5
		3	2.1	LAAEKEEEQDEEEVSDSKVAVEE	11.0	0.58	12,94
		4	1.8	CLVLAAFYLLAFLLFLLLMALLL	19.0	0.87	12,90
		5	1.8	QEIERLLKRAEQCAEKQRMEEEE	8.2	0.44	2,15
α_1	2.0	6	2.1	KEEREAAKQTKKRERDQEDKEKA	6.7	0.83	1,9
		7	1.9	RENLKEATARDAKMEKASWHKTA	15.3	0.38	12,97
		8	1.7	ALYAVKIVLLFMLVFAMLLAIIF	17.6	1.15	12,97
		9	1.8	LAAAAFAKAARGVNAKCLGASKC	7.1	0.91	11,83
		10	2.0	QEQEKRHKENDEADAEESKRAER	5.7	0.92	1,7
		11	1.9	AWYTVRWWTILLMAHLLVMVLLL	24.5	0.90	12,97
		12	2.0	YLAFLEFMFTLITVHFEALYEHT	14.7	0.43	11,83
		14	2.4	GRQESETWTDKTEEHDGSGKEYY	6.2	-0.46	11,84
β_2	2.1	15	2.3	VAVAEEVIIRVACMAYCQGTSIA	9.7	0.21	11,83
		16	1.7	NLLAITVITALTVVLIVLVYTVV	7.4	0.31	11,83
		18	1.5	YIRPSVPADSGVDNTHPAGGAVS	9.5	0.57	3,22
		19	1.5	SSGSCGSGNTGGSSNQCCSGSGS	2.9	1.60	3,22
		20	1.5	GGAGAGVVVGGGGGGGGGASGVA	7.1	1.91	9,69
		21	1.4	SFDTTKIVTARTSTTRASSSMTT	9.0	0.97	3,24
		22	1.2	PAEMEQYPTEGGVEHGACENCQA	9.5	0.83	3,18
		23	1.3	AAGGSAMAGSQAAAAAASAAALG	12.6	1.92	10,75
		24	1.2	NNINILVVCNFLVCLEVNTNSIM	12.7	0.68	6,49
α_3	1.5	25	1.3	FQEQNFSFDEAREMENAELNTHH	9.3	0.50	3,27
		26	1.4	AALVWNENIAAQVALINNVTVGI	11.4	0.75	11,82
		27	1.4	VVLSILLAVAVAAVALSAAVMLA	19.4	1.38	9,65
		28	1.4	YIIILFIVLIPLMLLFIFLTLVL	21.3	1.35	4,35
		29	1.8	TARMKQRRWRLEKRQPLKLSLIM	8.3	0.62	3,26
		30	2.1	AAAATAQEGMNLMLVVATAAASA	5.4	0.54	3,27
		31	1.9	LMFHVIWFLAIAAMMLFIALLLL	24.5	0.96	9,63
		32	2.3	VMCTALSTDKLRLRGLIFRVVSG	12.0	0.46	4,35
		34	2.2	RRAETEGGEEARRKEPHSVLHKE	10.3	0.52	5,40
		35	1.9	IIIVIVVVIIDVFIVVVIAVIVV	21.3	1.90	3,28
β_4	1.9	36	1.7	MALIIIVILVYFLIIFLLYLVVI	18.5	1.82	6,48
		37	1.7	GAYLTVIILGVVTKVIMSAFTAT	15.9	0.61	6,49
		38	1.7	LDCETLHITFDSFFPSTFSDTHP	10.3	0.40	5,42
		40	2.5	MRRHKVVIRFVQLLVASPKVTGV	14.3	0.45	4,36
β_5	2.0	41	1.8	PRTPYVMIFAFTFTPIWIAYVNM	11.8	0.42	4,37
		42	1.9	YFVVVEPEFIDYATVNFAILIAV	17.3	0.89	4,36
		43	1.7	KAPPPVVCPPARPPDIDPPQDAD	3.3	0.54	4,38
		45	2.7	PQVTFTCNRLAGIITITVPRVQT	13.0	0.56	4,36
		46	2.1	KQKSLKKKLAYDAAKVKAFLNYK	12.0	0.70	3,32
		47	1.7	LMVLVAIAVAMMAGVAIGFALIA	12.8	1.13	9,65
β_6	1.7	48	1.3	IVVISLMLIFHVVVIIVIVVFEI	18.5	1.50	4,36
		49	1.3	IVYAIVLLLIYLIIMVFIVIFMI	16.9	1.55	9,65
		50	1.6	ATVIMVLFSACLILPSLVAQTVP	17.0	0.98	8,58
		51	1.3	GECEANCCHEPHEEVTEETLETV	7.8	0.75	9,67
		53	2.8	SGGGGGGGGIIGMIGGMHCGLGG	8.9	1.00	15,122
T_7	2.7	54	2.5	RDQRTAKRQEAAYVLQEASTRLA	0.0	0.40	-
		55	2.7	NSLVIVVVVLINHPMPVSFICLP	1.0	0.61	6,51
		56	2.7	LAIVQYWWLPDLPDAVHDDYVRA	12.0	0.86	6,50

L_8	2.4	58	1.8	LIFIVLMIILFIILLVILVVVAI	17.0	1.78	6,50
		59	2.5	RQRCTAAKMKGARRAYVRGVSVY	17.9	0.65	9,64
		60	3.0	KQTTRRDDARASKEKEADPKKIA	0.0	0.46	-
β_9	1.5	62	2.0	AGAITFVLVMVAIAVIACIDACI	13.2	1.09	6,48
		63	1.5	YWINLYRMRLVLLLPFIAWYLKP	24.5	1.43	3,31
		64	1.0	LLVVVLVLVLVPLLLVIFLIVSV	17.9	1.63	8,59
		65	0.8	MMVHHVIWLIALIVLVMLHECII	19.1	1.40	3,27
		66	0.9	AVAIVSSSVLAIAFQMIVVFIYE	13.0	0.93	10,74
		67	1.1	DDDDDDDDDDDDDDDDDDDDDDD	7.9	3.27	6,51
		68	1.4	MDEGAEEEGEFLEEASNEASGMA	13.4	0.85	10,74
		69	1.7	AAAAVIIIAASAIVAATVAATVA	10.4	1.43	10,77
		70	2.2	HHYRQYHHQQSYAMHRWQYWFYH	10.9	1.20	15,122
		71	2.3	IGIITEMFSTTQTTAFATAVAYA	11.3	0.80	9,68
β_{10}	1.4	73	1.0	VLASYLAIYIVLILIAVVSILIT	18.3	1.41	3,27
		74	1.1	VLIVATLTYLIILTALSMFVVLA	15.0	1.37	11,83
		75	1.1	TVLMSVLFATYACTTTITNVLLI	12.8	0.93	11,82
		76	1.5	TVRFFNTTFLAALLFMQFFQHFF	13.3	0.56	1,12
		77	2.1	TTTCSGSATSGSGGSSASNSSTS	10.5	1.55	9,69
β_{11}	1.5	79	1.5	AGCGVITGVLLTLVGGITCTAFG	14.2	0.70	12,93
		80	1.7	GAGSGGGMGATGSGGGGGSSGGG	6.7	1.90	10,77
		81	1.4	MAFVIYASGAICAAALTGAQCWI	17.5	1.95	1,8
		82	1.4	IVTVTAYNLIVLTYVVAFMIIAV	12.8	1.37	10,75
		83	1.5	FLLVYCGIYVILLGVAVFWHSLV	15.0	0.75	10,74
		84	1.5	YCAGIGIIITVATGFICHVKGIT	18.2	0.60	10,73
α_{12}	1.9	86	1.9	HPSANDSTADPPNNNADDRFNPP	7.2	0.61	12,89
		87	1.6	NPTGVTVTYPTFPPEGAPRYPRD	8.0	0.40	12,90
		88	1.7	HAPVPILLPLFHLLFRDSFPNEL	9.2	0.59	12,91
		89	1.5	AQVLGAAGQPAGAASDTKRMATN	12.2	0.65	11,79
		90	1.5	IAAAIQLYYAWACMAMANSYAQA	19.0	0.93	1,4
		91	1.7	AQDAVWTKAAYKATSAYLLAYLA	9.6	0.92	12,94
		92	1.8	GAIAGAAAGVLIAAVAILKALRI	8.0	1.02	12,89
		93	1.7	VLAGATHCLGSVAGLMTLMLISA	14.2	0.67	1,4
		94	2.0	ARAIALIELLGSNIRVSAWDILL	14.7	0.52	1,3
		95	2.4	VAQIREAIARLTAKAQREFVRKA	7.6	0.74	12,92
		96	2.4	ASAAAAACAVVVSTQGGVVNGVQ	8.6	1.03	12,93
		97	2.5	LLLLMLYYLLFLLLLLIFAMLL	24.5	1.70	1,11
α_{13}	2.2	99	3.2	VAEDEMAGEAEAQGEEVSAKQFA	8.4	0.81	13,102
		100	2.6	YRRRAALCRRMVQTQYRNYLQGL	7.0	0.52	13,103
		101	2.2	QRVLARRIYADRVYILLAIWQFN	10.5	0.50	13,104
		102	2.6	HEAKKEIKQNKSAEAHRARHNKA	8.4	0.83	13,99
		103	2.5	QKQEHAYVAVIGDYVYQHKDSKR	7.0	0.72	13,100
		104	2.1	VLIDMYLIIMNMILRRHAHLTLR	10.5	0.60	13,101
		105	1.9	VAIHERKDCGQLEDDIHGVLARA	7.6	0.50	13,108
		106	2.3	AAAEKRDKQDQRVQEGEKQILEA	6.4	0.74	13,109
		107	2.2	NLENLRNNRRKLQIRQSVLTRRI	10.8	0.57	17,136
		108	1.7	CIRARRLQSLARLTWVSLSTMWA	14.7	0.70	17,136
		109	1.9	RTERSDTREREERKTKLLHIAVA	8.3	0.67	13,112
		110	2.2	ARYFKLYIFRLQERLYKTEEENQ	7.2	0.53	13,107
		111	1.9	LFLLLLIVLGLLQLLLVGFAIIY	18.4	1.04	17,134
		112	1.7	SRIAVLAKWLYALSSGALERLTH	15.2	0.63	14,116
		113	2.0	EAALSEDDRLGGADRDELAKEAQ	6.0	0.76	13,110
β_{14}	2.0	115	1.8	LIYMTYAQLLNFIMAALFLKAYL	13.4	1.07	17,133
		116	1.8	IQIVLVWWVLVSGFMVFCVFVAF	15.2	1.15	13,112
		117	1.7	LPLISLLISGPRVGFFSSCMSWI	13.3	0.41	17,133

		118	1.8	VLALVMDDFVFLVFRVFFFLFVI	16.5	1.20	17,134
		119	2.0	DIRRSDLFTEQGEFIDVDRTEKR	6.1	0.52	15,122
		120	2.9	LVFTFTRRVILATFPCLTLTVCV	13.8	1.16	17,132
		122	2.9	EAFIQAKEVTDKMHVILILATFM	14.2	0.56	17,132
		123	2.8	KEKENEAFQRKEARDNASENAQA	6.6	0.60	15,126
α_{15}	2.9	124	2.8	VKSALRLMKELEAGRANIRYKNS	0.0	0.57	-
		125	2.9	LLLLLLIILCFFLMLLYAILFGL	17.1	1.55	16,128
		126	3.1	ERWRSLEHEMLGQLVYLRNRVRK	6.6	0.00	15,123
		128	2.0	CVLAVVIFLMLYRLAAILLPDTT	20.8	0.84	17,134
β_{16}	2.0	129	2.0	NTRNSVMDRNKMPAFVLGVEQQG	9.0	0.41	17,133
		130	2.1	KADATPPETIGVFPAEAGPKPNL	11.9	0.62	17,132
		132	2.0	LLLIIVVEIFMSLFVMIILIVVC	14.9	1.75	16,128
		133	1.7	RRRRRRRRRRRRIEWRRRRLRRS	13.4	2.00	14,115
β_{17}	1.9	134	1.8	LLILILLIAIAIYALLLFFFFLL	20.8	1.43	16,128
		135	1.8	GTTVSSNNCASNLGHTHRALSSP	11.4	0.76	14,115
		136	1.8	TLVTLYALCPIIMFHIIPIMFLL	14.7	0.95	13,108
		137	2.4	PTGHCAGAHPYAPTRAGTCTGIF	9.5	0.67	13,104

Insieme a queste posizioni, altri siti non coinvolti direttamente in alcuna interazione con il cofattore od i substrati mostrano un alto grado di conservazione di sequenza, comparabile alla conservazione osservata nei residui funzionalmente importanti (*R* ≥ 1.0; Tab. 5.II). Questi siti possono essere raggruppati in due categorie distinte: 1) posizioni ricche in glicina/alanina; 2) posizioni occupate principalmente da residui idrofobici (il sito 97, ad esempio, che mostra un punteggio di conservazione evolutiva di 1.7 deviazioni standard dal valore medio di conservazione, è quasi invariabilmente occupato da una leucina od un residuo aromatico in tutte le 921 sequenze considerate, sebbene non sembri implicato in alcun ruolo funzionale).

Le posizioni occupate principalmente da glicina od alanina (23, 80 e 92), che mostrano un alto grado di conservazione (1.9, 1.9 e 1.0, rispettivamente), possono giocare ruoli importanti al di là di quelli collegati al legame del PLP od alla formazione di contatti idrofobici. Per esempio, due siti ricchi in

alanina (23 e 92) sono presenti al centro di una α-elica: è stato osservato che l'alanina mostra il maggior grado di propensione rispetto a qualsiasi altro residuo per tale posizione (Richardson & Richardson, 1989). Sembra che questa preferenza sia dovuta alle particolari caratteristiche strutturali dell'alanina, che dirigono e stabilizzano il ripiegamento dell'α-elica (Blaber *et al.*, 1993). L'altro sito, ricco in glicina (80), è stato trovato nella SCR β_{11}, dove può essere responsabile della curvatura del foglietto (Richardson & Richardson, 1989).

5.3 Contatti idrofobici conservati

Per verificare l'ipotesi che la conservazione delle proprietà chimico-fisiche del secondo gruppo di posizioni sia la conseguenza della pressione selettiva esercitata su questi siti per mantenere la stabilità degli enzimi PLP-dipendenti di tipo I, attraverso il coinvolgimento dei residui corrispondenti in interazioni idrofobiche, è stata condotta un'analisi dei contatti idrofobici conservati (CHCs) sulle SCRs precedentemente identificate.

Studi precedenti, condotti a livello comparativo, e focalizzati sulla relazione tra conservazione di sequenza di una famiglia proteica ed i contatti idrofobici delle corrispondenti strutture disponibili (Ptitsyn, 1998; Hill *et al.*, 2002; Gunasekaran *et al.*, 2004; Gromiha *et al.*, 2004) hanno considerato due residui in contatto se la distanza tra i loro carboni α e/o tra qualsiasi atomo si trovava al di sotto di un limite prefissato arbitrario. In questo lavoro è stato adottato un criterio differente, basato sull'analisi comparativa dell'area di contatto apolare per ogni possibile coppia di residui appartenenti alle SCRs (Drabløs *et al.*, 1999).

I CHCs sono quindi definiti come contatti idrofobici tra residui, che coinvolgono solo atomi apolari (Drabløs, 1999), osservati in almeno due delle strutture analizzate. Questo approccio ha permesso di quantificare la "forza" di un contatto idrofobico e di stabilire la correlazione tra questa grandezza e la conservazione evolutiva dei siti corrispondenti. I CHCs più intensi per ogni sito, appartenenti alle SCRs, e le posizioni corrispondenti coinvolte nell'interazione idrofobica sono mostrati nella Tabella 5.II.

La figura 5.6 mostra i valori di conservazione evolutiva media tra coppie di siti coinvolti nei CHCs, in funzione dei valori medi di contatto idrofobico. I residui che interagiscono con il PLP, così come i siti ricchi in alanina/glicina suddetti, non sono stati riportati nel grafico, dal momento che la loro alta conservazione evolutiva riflette probabilmente altre funzioni rispetto alla stabilizzazione della struttura dei membri di questa superfamiglia attraverso la formazione di contatti idrofobici. Da questa analisi è risultato un coefficiente di correlazione significativo (r = 0.70) tra le due variabili. La significatività statistica di r è stata stabilita con un *t-test*, assumendo $r = 0$ come ipotesi nulla. Il risultato (*p-value* $\cong 1.7 \times 10^{-53}$) indica che c' è una correlazione statisticamente significativa tra la forza di un CHC ed il grado di conservazione evolutiva dei residui coinvolti.

A valori > 16 $Å^2$, il grado di conservazione medio diventa comparabile ai valori misurati per i residui importanti da un punto di vista catalitico ($R \geq 1.0$). I CHCs con i più alti valori di area di contatto media apolare (Tabella 5.III) possono essere raggruppati in tre insiemi principali (Fig. 5.7): un primo insieme è localizzato in una regione nascosta al solvente alla base del nucleo comune conservato del dominio maggiore legante il PLP, formato da sei SCRs (α_3 β_6, L_8, β_9, β_{10} and β_{11}); un secondo, piccolo insieme, si trova attorno alla posizione 133 del dominio minore (α_{13}, β_{14}, α_{15}, β_{16} and β_{17}); un

terzo insieme di CHCs forma una cerniera tra le SCRs α_1 e α_{12}, poste rispettivamente all'inizio ed alla fine del dominio maggiore.

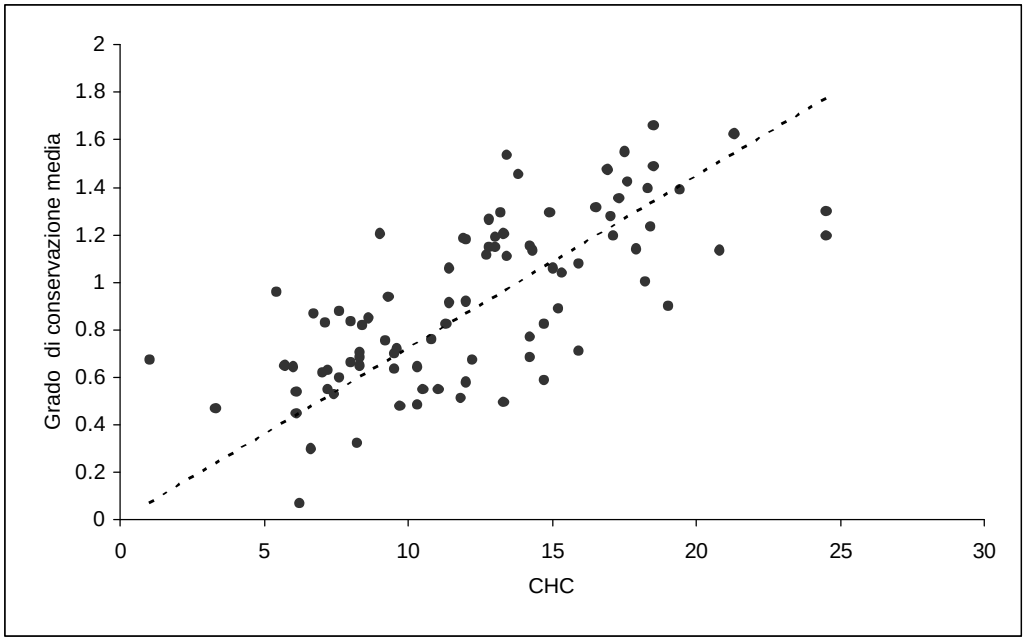


Fig. 5.6 Conservazione evolutiva media tra coppie di siti coinvolti nei CHCs, in funzione dei valori medi di contatto idrofobico

Tabella 5.III. CHCs con i più alti valori di area di contatto media apolare.

		Contatti idrfobici (Å^2)															
	I SCR, sito	**1**, 4	**1**, 8	**1**, 8	**1**, 11	**3**, 27	**3**, 27	**3**, 28	**3**, 31	**4**, 36	**4**, 36	**6**, 49	**6**, 50	**8**, 59	**10**, 73	**13**, 111	**15**, 125	**16**, 128
	II SCR, sito	**12**, 90	**11**, 81	**12**, 97	**12**, 97	**9**, 65	**10**, 73	**4**, 35	**9**, 63	**6**, 48	**5**, 42	**9**, 65	**8**, 58	**9**, 64	**11**, 84	**17**, 134	**16**, 128	**17**, 132
Molecola (Codice PDB)	**1BJ4**	18.5	17.7	17.5	16.3	17.9	19.7	26.0	32.1	17.2	20.2	20.2	23.0	29.9	22.4	25.9	20.0	23.6
	1BS0	18.2	22.4	18.5	32.7	22.5	24.5	27.1	29.1	15.7	9.9	24.2	17.8	16.7	18.2	21.0	27.6	27.6
	1FG3	17.4	43.0	26.0	15.7	22.2	15.3	29.7	34.4	21.5	23.9	18.2	27.0	-	19.7	24.6	30.8	39.1
	1JG8	19.0	18.7	11.8	16.0	12.9	2.8	21.7	5.7	31.1	22.6	5.5	37.4	21.0	12.4	32.1	11.9	20.3
	1ECX	17.3	28.8	22.8	23.4	16.7	34.5	10.6	16.5	-	23.6	-	-	20.9	11.9	29.4	26.6	22.2
	1BJW	-	12.2	20.0	-	24.3	41.5	18.5	27.3	22.7	-	17.2	25.8	24.2	15.9	27.4	27.5	25.2
	1D2F	38.5	-	17.8	40.0	23.0	20.0	29.3	27.4	19.6	-	20.7	15.7	7.4	-	25.7	-	32.8
	1C7N	32.6	-	18.1	44.5	20.3	13.2	16.7	32.6	20.5	-	28.0	26.6	-	-	23.4	-	29.5
	1ELQ	28.7	3.2	21.9	8.5	23.3	39.5	17.9	3.6	-	20.9	-	-	31.0	34.6	19.1	28.8	25.0
	1DGD	18.2	19.4	16.4	29.3	17.7	19.3	32.5	26.7	22.8	19.5	27.2	12.2	37.9	10.6	-	30.0	27.7
	1BJN	10.8	31.5	31.0	27.8	13.8	17.4	-	17.3	19.4	-	-	-	-	24.9	15.6	23.3	21.5
	1AY4	21.7	22.5	8.2	31.3	22.4	12.0	19.0	15.4	22.6	18.7	23.1	23.6	-	18.4	19.2	-	-
	1DTY	26.8	20.8	16.7	33.4	19.0	13.7	42.0	17.8	22.4	13.4	29.3	23.8	6.9	10.1	-	-	-
	2GSA	29.7	2.8	14.7	10.6	19.7	26.7	23.9	20.7	24.0	20.0	20.3	24.3	18.5	10.5	-	31.4	-
	1B9H	14.3	17.3	20.3	13.2	19.1	13.0	28.1	20.8	34.0	29.3	17.5	24.0	14.8	33.2	20.4	18.4	21.9
	1AX4	26.6	12.9	2.2	23.0	18.1	12.4	-	37.3	-	-	15.3	-	32.5	21.4	26.2	9.5	19.2
	1CL1	19.4	17.5	17.4	33.7	12.2	6.6	20.9	37.5	21.5	24.2	19.0	31.5	32.8	27.7	14.7	11.4	25.0
	1GTX	1.8	-	3.3	30.9	19.9	14.2	32.5	20.9	28.1	17.7	25.2	15.0	31.0	22.8	-	26.2	31.3
	1JS6	11.8	14.6	22.8	39.9	15.6	2.8	13.7	31.4	21.7	31.8	18.5	17.9	21.8	20.0	52.0	28.2	27.1
	1ORD	12.4	0.8	13.8	12.1	14.9	25.2	0.4	37.0	-	34.4	-	-	22.1	18.6	-	23.7	22.0
	1QGN	18.6	27.4	23.8	25.4	20.0	12.2	25.0	22.0	19.9	22.1	18.0	0.2	14.2	14.0	-	-	-
	1LK9	16.2	40.1	15.3	28.8	23.0	29.9	33.0	21.9	12.2	16.0	28.6	15.4	-	23.5	22.9	-	21.4
	1MDX	18.9	29.1	24.5	28.3	21.0	5.2	20.6	27.6	27.7	29.2	12.2	29.0	28.6	12.9	24.3	18.7	16.0
	Media	**19.0**	**17.5**	**17.6**	**24.6**	**19.1**	**18.3**	**21.2**	**24.5**	**18.4**	**17.3**	**16.9**	**17.0**	**17.9**	**18.2**	**18.4**	**17.1**	**20.8**

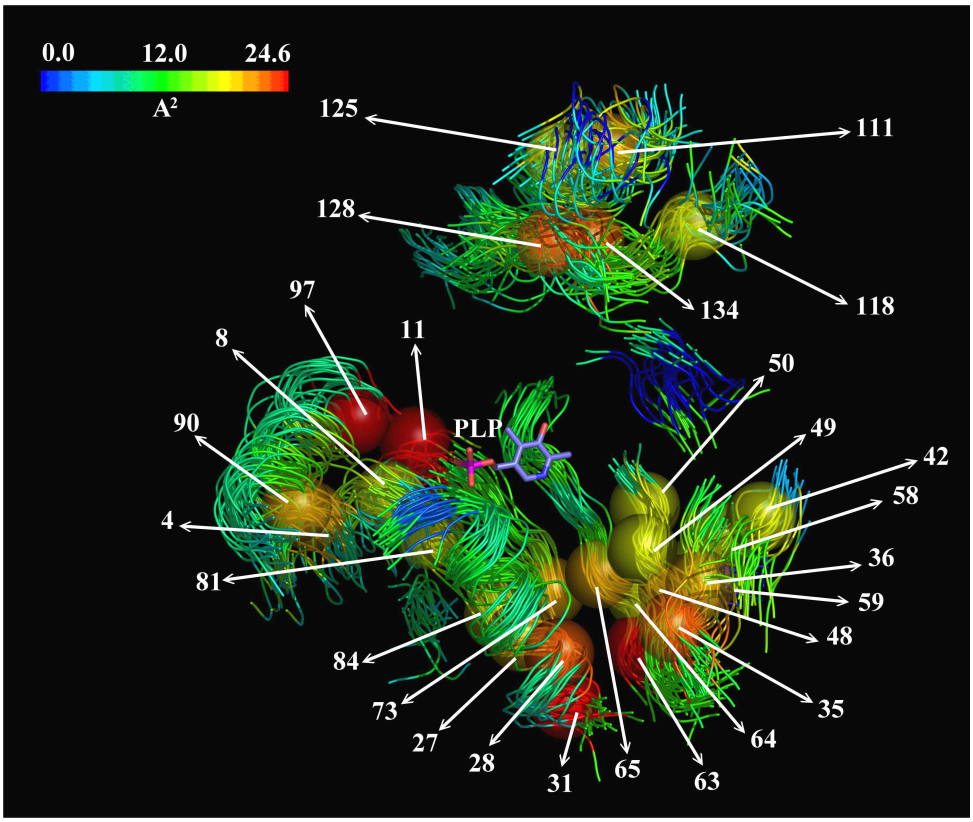


Fig. 5.7 Rappresentazione dei siti coinvolti in contatti idrofobici conservati (CHCs). Le posizioni coinvolte nella formazione dei CHCs con superficie apolare media di contatto > 16 Å^2 sono rappresentate come sfere colorate, ed etichettate in base alla numerazione indicata nell'allineamento mostrato in Fig. 5.2. Le catene polipeptidiche delle 23 strutture sovrapposte sono mostrate come nastri e colorate in accordo con il valore medio di contatto idrofobico. Il PLP è mostrato in azzurro, con gli atomi di ossigeno in rosso, l'azoto blu, ed il fosforo viola.

Gli amminoacidi che appartengono al primo insieme di CHCs si trovano in posizione 27, 28 e 31 in α_3; 35 e 36 in β_4; 42 in β_5; 48, 49 e 50 in β_6; 58 e 59 in L_8; 63, 64 e 65 in β_9; 73 in β_{10}; 81 e 84 in β_{11} (Fig. 5.7; Tab. 5.II). I cinque residui che partecipano alla formazione del secondo insieme (111 in α_{13}, 118 in β_{14}, 125 in α_{15}, 128 in β_{16} e 134 in β_{17}), sono localizzati in prossimità della posizione 133, di β_{17}, che è occupata principalmente da un residuo di arginina

(18 delle 23 strutture analizzate): il gruppo carbossilico di molti substrati legati dagli enzimi di tipo I forma una coppia ionica con questo residuo (Jansonius, 1998). I residui che formano il terzo gruppo di CHCs sono coinvolti in contatti inter-elica in 22 delle 23 strutture considerate (la sola eccezione è rappresentata da 1BJW, nella quale solo due CHCs che coinvolgono il residuo 8 sono conservati; Tab. 5.III). Questi residui (posizioni 4, 8 ed 11 della SCR α_1; 90 e 97 della SCR α_{12}), localizzati ai lati delle due eliche che delimitano il dominio maggiore, occupano posizioni *i*, *i* + 3 ed *i* + 7. Il sito 97, che è stato descritto in precedenza (Paragrafo 5.2; *R* = 1.7) è impegnato nella costituzione dei CHCs più estesi che formano la cerniera tra le SCRs α_1 ed α_{12} (posizioni 11-97 ed 8-97, Fig. 5.7; la superficie di contatto apolare media è rispettivamente di 24.6 $Å^2$ e 17.6 $Å^2$), ed in un ulteriore contatto conservato con la posizione 7 (15.3 $Å^2$).

5.4 Implementazione degli algoritmi utilizzati per l'analisi nei programmi CAMPO, SCR_FIND e CHC_FIND, e loro convalida

Parallelamente allo studio focalizzato sulle regioni strutturali, i contatti idrofobici, ed i residui evolutivamente conservati della superfamiglia degli enzimi PLP-dipendenti di tipo I, e nel tentativo di estendere questo tipo di analisi ad altre superfamiglie proteiche, gli algoritmi sviluppati per l'indagine *ad hoc* sono stati generalizzati in programmi resi disponibili alla comunità scientifica, e la loro efficacia predittiva valutata su sistemi proteici noti e ben caratterizzati.

Negli ultimi 30 anni sono stati proposti numerosi metodi per quantificare la conservazione evolutiva di un residuo, ma, sino ad oggi, non esiste ancora alcun criterio matematico per valutare oggettivamente l'efficacia di algoritmi

preposti a tale compito. Recentemente, Valdar (2002) ha confrontato diversi approcci differenti per quantificare la conservazione evolutiva in una data posizione di un allineamento multiplo: nessun metodo sembra raggiungere allo stesso tempo rigore statistico e biologico, ed in molti casi il risultato ottenuto è fortemente dipendente dalla famiglia utilizzata nell'analisi (numero di sequenze, metodo di allineamento, scelta della matrice di scambio *etc*.).

Alla luce dell'analisi compiuta da Valdar, il principale obiettivo nell'implementazione dell'algoritmo utilizzato per l'analisi dei ECRs degli enzimi PLP-dipendenti di tipo I, in un programma da rendere disponibile alla comunità scientifica, è stato quello di dare all'utente la possibilità di avere pieno controllo sulla scelta dei parametri utilizzati per costruire l'allineamento multiplo ed assegnare un punteggio di conservazione in ogni posizione. Infatti, come discusso da Valdar (2002) ed Armon (2001), questi due passaggi sono principalmente responsabili dell'accuratezza nella valutazione della conservazione evolutiva di una famiglia o superfamiglia di proteine.

CAMPO (http://schubert.bio.uniroma1.it/CAMPO; Fig. 5.8) fa uso di una procedura simile a quella adottata da ConSurf (Glaser *et al.*, 2003), uno dei programmi più utilizzati nell'analisi di regioni evolutivamente conservate, per ottenere in maniera automatizzata un allineamento multiplo a partire da una sequenza sonda.

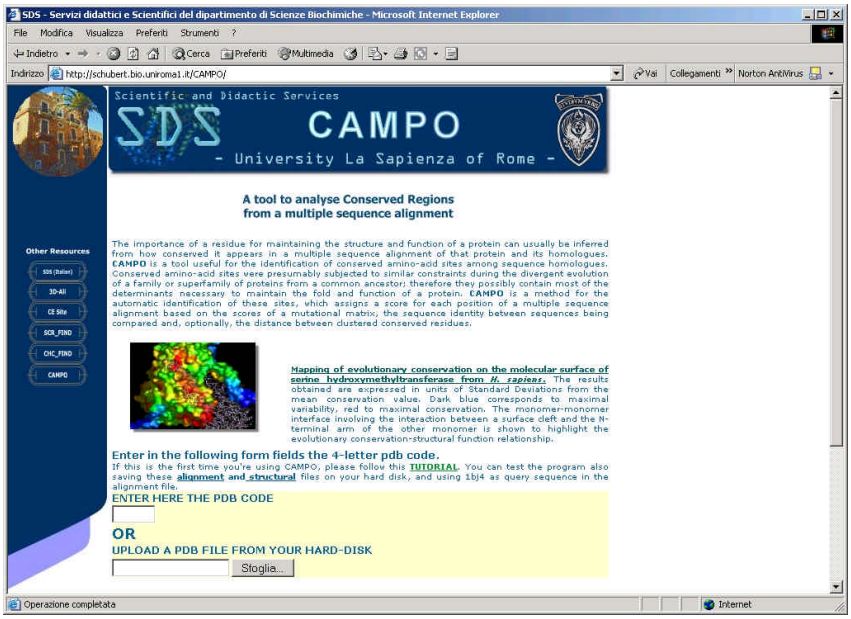


Fig. 5.8. La pagina iniziale di CAMPO

CAMPO utilizza il programma Blast (Altschul *et al.*, 1997), installato localmente, per compiere la ricerca di proteine omologhe alla sequenza sonda, con una soglia di significatività necessaria per accettare o scartare le sequenze decisa dall'utente. Le proteine selezionate sono ulteriormente filtrate (vedi sotto) e successivamente allineate utilizzando il programma CLUSTALW (Thompson *et al.*, 1994); inoltre CAMPO permette la scelta dei seguenti parametri: 1) banca dati proteica utilizzata per collezionare sequenze omologhe (NRDB [Holm & Sander, 1998], PDB [Berman *et al.*, 2000] e SWISS-PROT [Junker *et al.*, 2000] sono attualmente selezionabili, ed altre banche dati possono essere facilmente aggiunte); 2) minima e massima percentuale di identità di sequenza, e minima percentuale di residui allineati alla proteina sonda per accettare o scartare le sequenze trovate. Infine, CAMPO permette all'utente di scegliere la matrice di mutazione più appropriata (attualmente sono disponibili le serie PAM e BLOsum) per

allineare ed assegnare un punteggio di conservazione alle sequenze filtrate (Fig. 5.9).

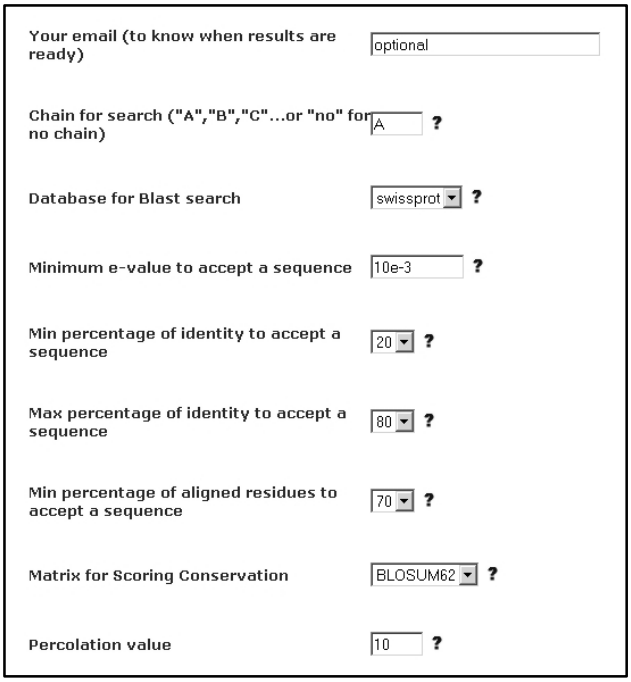


Fig. 5.9. I parametri selezionabili dall'utente di CAMPO

Rispetto a CAMPO, ConSurf utilizza esclusivamente parametri fissati per costruire un allineamento multiplo delle sequenze trovate, ed una matrice di scambi amminoacidici, derivata da Miyata *et al.* (1979), basata sulle caratteristiche chimico-fisiche tra gli amminoacidi, per quantificare le sostituzioni tra residui. Dal momento che è stato dimostrato che l'utilizzo di matrici basate sugli frequenza degli scambi amminoacidici osservati in natura, come le PAM o le BLOsum, è solitamente preferibile a quello di matrici basate su proprietà chimico-fisiche dei residui (Vogt *et al.*, 1995), è sembrato appropriato utilizzare queste ultime per assegnare un punteggio agli scambi.

L'algoritmo implementato assegna un punteggio ad ogni colonna dell'allineamento multiplo di sequenza attraverso l'applicazione di una matrice PAM o BLOsum, ed incorpora un sistema di pesatura basato sulla similarità di sequenza delle proteine confrontate (Paragrafo 4.5). I risultati possono essere utilizzati per colorare l'allineamento e la superficie di una struttura proteica di riferimento, per permettere la facile identificazione di regioni funzionalmente importanti (Fig. 5.10).

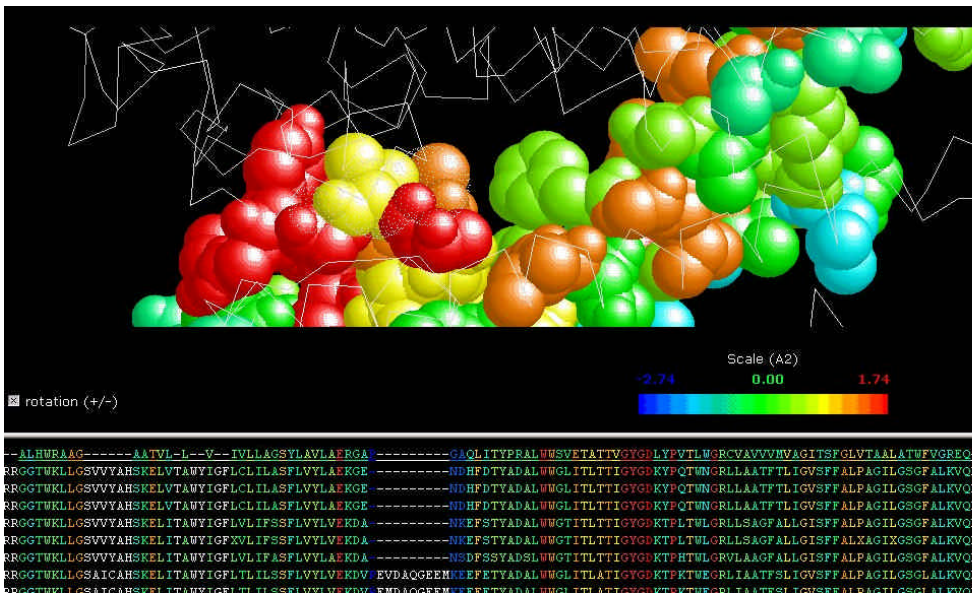


Fig. 5.10. Un esempio di risultati ottenuti con CAMPO con la proteina 1BL8 (canale del potassio da *Streptomyces Lividans*)

In maniera opzionale, CAMPO permette di misurare la conservazione evolutiva di una regione proteica di raggio arbitrario, centrata su ogni atomo della struttura tridimensionale, applicando una tecnica basata sulla teoria della percolazione (Harrison, 2001). La teoria della percolazione è una teoria statistica che indaga la formazione di insiemi e l'esistenza di fenomeni di percolazione delle proprietà di un sistema. La teoria suppone l'esistenza di

una griglia regolare alla base del sistema. Un insieme viene definito come un gruppo di siti vicini occupanti la griglia, ed il limite al di sotto del quale una proprietà del sistema tende a diffondersi, percolando, (la conservazione evolutiva, nel nostro caso), è detto "limite di percolazione". Questo approccio permette all'utente di raggiungere una migliore risoluzione delle regioni conservate sulla struttura tridimensionale, possibilmente responsabili dell'interazione con piccoli ligandi e/o macromolecole.

Per dimostrare la capacità di CAMPO di individuare ECRs necessarie probabilmente all'attività ed alla stabilità proteica, viene di seguito riportata come esempio l'analisi condotta sul canale del potassio (K^+) di *Streptomyces lividans* (Kcsa, PDB: 1BL8), e la comparazione di questi risultati con quelli ottenuti utilizzando ConSurf. Kcsa, una proteina ben caratterizzata e per la quale è disponibile molta informazione strutturale e di sequenza, è un canale integrale di membrana molto simile a tutti i canali del potassio conosciuti, in particolare nella regione del poro (Doyle *et al.*, 1998). E' stato osservato che la conservazione di sequenza tra i canali del K^+ è maggiore per i residui che costituiscono la regione del poro (residui 61-85) e l'elica interna (86-119), mentre l'elica esterna, N-terminale (23-60), è meno conservata (Doyle *et al.*, 1998).

I risultati ottenuti da CAMPO utilizzando la struttura corrispondente al codice PDB 1BL8, catena A, sono rappresentati in Fig. 5.11. CAMPO ha identificato 68 sequenze omologhe utilizzando i parametri preimpostati (limite di significatività di Blast, *E-value* 0.001; minima e massima percentuale d' identità per accettare una sequenza: 20% ed 80%, rispettivamente; minima percentuale di residui allineati alla sonda per filtrare le sequenze trovate: 80%).

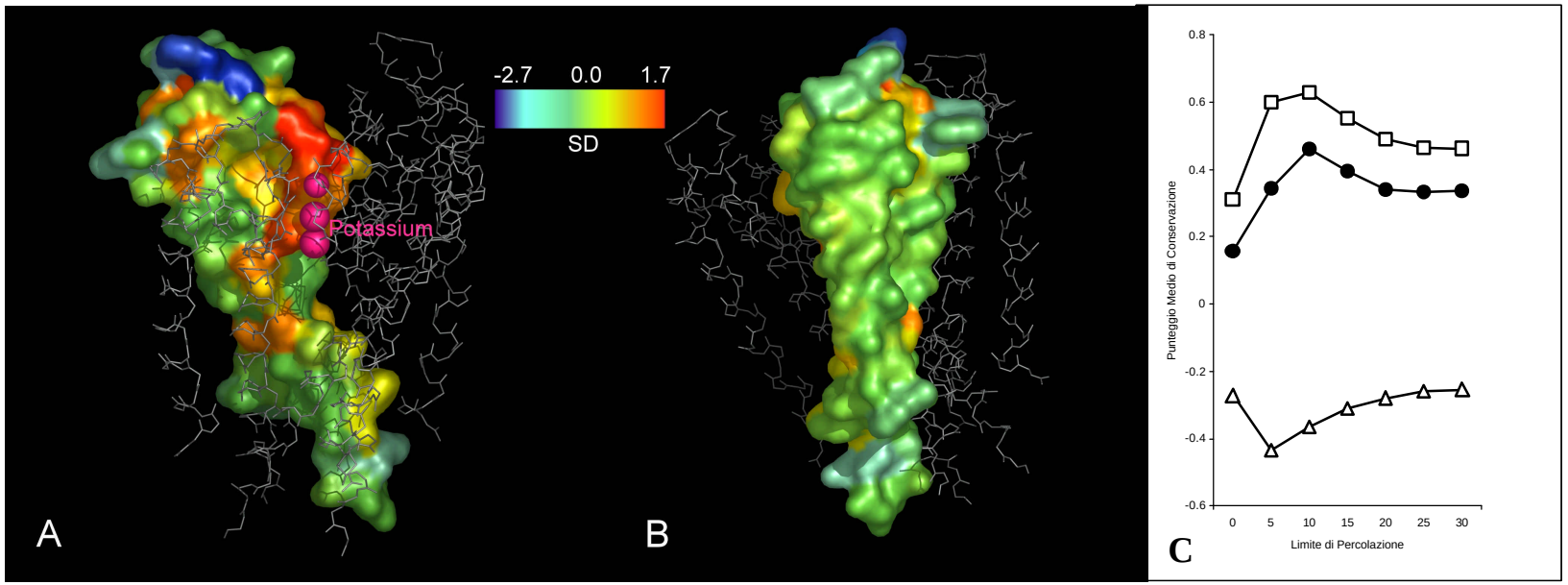


Fig. 5.11. Conservazione evolutiva della superficie (A) interna ed (B) esterna del canale del potassio da *S. lividans*, in base ai punteggi registrati da CAMPO, e (C) grafico dei punteggi medi di conservazione evolutiva in funzione del limite di percolazione per l'elica esterna (triangoli), l'elica interna (cerchi pieni) e la regione del poro (quadrati) del canale del potassio. I risultati ottenuti sono espressi in unità di deviazioni standard (SD) dal valore medio di conservazione. Secondo lo schema di colorazione utilizzato da CAMPO, il blu scuro corrisponde alla massima variabilità, il rosso alla massima conservazione. Gli ioni potassio sono mostrati come sfere rosa.

Una matrice BLOsum62 è stata scelta per allineare le sequenze ed assegnare un punteggio di conservazione. CAMPO è stato in grado di individuare i residui maggiormente conservati che si affacciano sulla superficie interna del canale (Phe 114, Leu 110, Val 106, Gly 104, Leu 105, Gly 99, Ile 100, Thr 74, Thr 75, Trp 68 e Pro 83) e quelli che interagiscono con le altre subunità che formano la struttura tetramerica (Trp 67, Tyr 78 e Asp 80). In particolare, i residui Gly 77, Tyr 78, e Gly 79, che interagiscono con lo ione K^+ e che sono assolutamente indispensabili per la selettività ionica del canale, sono stati indicati come i più conservati all'interno del canale. La differenza tra la superficie interna ed esterna del canale è resa ancora più evidente applicando la percolazione della conservazione evolutiva per individuare le regioni maggiormente conservate (Fig. 5.11 A e B). E' stato adottato un limite di percolazione di 5 Å, nel tentativo di massimizzare il rapporto tra il numero di atomi chiusi nella sfera effettivamente in contatto con l'atomo centrale, ed il numero di atomi non in contatto con esso. Inoltre, con un valore di 5 Å, le differenze tra i valori di conservazione media ottenuti per l'elica interna, l'elica esterna e la regione del poro sono i più evidenti (Fig. 5.11 C). Questo risultato è in accordo con quelli precedentemente ottenuti da Doyle *et al.*, 1998.

Risultati simili sono stati ottenuti dal programma ConSurf (si veda, ad esempio, http://consurf.tau.ac.il/gallery.html), che ha allineato 38 sequenze, compresa quella di Kcsa da *Streptomyces coelicolor*, che è identica al 100% alla sequenza sonda. Confrontando i punteggi assegnati dai due programmi, sulla base dello stesso allineamento generato da CAMPO, è stato ottenuto un coefficiente di correlazione *r* pari a 0.79 (Fig. 5.12 e Tab.5.IV).

Tabella 5.IV (pagine seguenti). Confronto dei punteggi di conservazione assegnati da CAMPO e ConSurf per 1BL8 (Kcsa da *Streptomyces lividans*)

Residuo	Punteggi assegnati da CAMPO	Punteggi assegnati da ConSurf (38 sequenze)*	Punteggi assegnati da ConSurf (69 sequenze)**	Punteggi assegnati da CAMPO (normalizzati)	Punteggi assegnati da ConSurf (normalizzati)
ALA23	-0.34	0.86	0.58	0.54	0.78
LEU24	-1.47	1.18	-0.40	0.28	0.50
HIS25	-0.21	0.61	0.69	0.56	0.81
TRP26	0.04	3.87	0.59	0.62	0.78
ARG27	0.08	-0.36	0.04	0.63	0.62
ALA28	-0.76	-0.39	0.21	0.44	0.66
ALA29	1.16	-0.51	0.34	0.87	0.70
GLY30	1.31	-0.27	0.58	0.90	0.78
ALA31	-1.31	-0.52	-0.27	0.32	0.53
ALA32	-0.17	-0.20	-0.21	0.57	0.55
THR33	1.34	-0.36	0.67	0.91	0.80
VAL34	0.36	-0.20	-0.23	0.69	0.54
LEU35	-1.3	-0.31	-0.35	0.32	0.51
LEU36	0.02	-0.38	0.82	0.62	0.85
VAL37	0.01	0.04	0.36	0.61	0.71
ILE38	0.04	-0.29	0.21	0.62	0.66
VAL39	-0.3	0.03	-0.03	0.54	0.59
LEU40	0.38	-0.29	0.23	0.70	0.67
LEU41	-0.09	-0.25	-0.25	0.59	0.54
ALA42	0.1	-0.01	-0.13	0.63	0.57
GLY43	-0.38	-0.07	0.20	0.53	0.66
SER44	1.1	-0.47	0.87	0.86	0.87
TYR45	-0.95	-0.21	-0.12	0.40	0.57
LEU46	-0.54	-0.14	0.10	0.49	0.63
ALA47	0.44	-0.28	0.09	0.71	0.63
VAL48	0.5	-0.07	-0.75	0.72	0.41
LEU49	-0.51	0.01	-0.63	0.50	0.44
ALA50	0.56	-0.39	0.47	0.74	0.74
GLU51	1.54	-0.55	1.02	0.96	0.92
ARG52	-0.73	-0.23	0.15	0.45	0.65
GLY53	-0.32	0.15	-0.75	0.54	0.41
ALA54	-1.25	1.07	-2.33	0.33	0.13
PRO55	-2.74	4.70	-4.32	0.00	0.00
GLY56	-2.48	4.70	-2.00	0.06	0.17
ALA57	-2.63	0.88	-2.31	0.02	0.13
GLN58	-1.1	1.28	-2.73	0.37	0.08
LEU59	1.12	-0.14	0.59	0.86	0.78
ILE60	-1.18	2.40	-2.17	0.35	0.15
THR61	0.48	-0.42	0.62	0.72	0.79
TYR62	0.23	-0.44	0.87	0.66	0.87
PRO63	0.09	-0.34	0.83	0.63	0.86
ARG64	-0.73	0.04	-0.32	0.45	0.52
ALA65	0.19	-0.43	0.42	0.65	0.73
LEU66	-0.29	-0.42	0.21	0.55	0.66
TRP67	1.45	-0.54	0.36	0.94	0.71
TRP68	1.32	-0.37	-0.12	0.91	0.57
SER69	-0.18	-0.47	0.38	0.57	0.71
VAL70	-0.26	0.05	-0.39	0.55	0.50
GLU71	1.24	-0.35	0.78	0.89	0.84
THR72	0.75	-0.62	0.85	0.78	0.86
ALA73	0.61	-0.37	0.54	0.75	0.76

THR74	1.04	-0.54	0.89	0.84	0.88
THR75	1.36	-0.66	0.97	0.92	0.91
VAL76	0.78	-0.51	0.62	0.79	0.79
GLY77	1.47	-0.66	0.94	0.94	0.89
TYR78	1.49	-0.66	0.87	0.94	0.87
GLY79	1.59	-0.66	1.08	0.97	0.94
ASP80	1.74	-0.66	1.24	1.00	1.00
LEU81	-0.5	-0.25	0.49	0.50	0.75
TYR82	-1.01	2.02	-2.74	0.39	0.08
PRO83	1.48	-0.66	0.91	0.94	0.88
VAL84	-1.22	1.56	-2.19	0.34	0.15
THR85	1.13	-0.58	0.93	0.86	0.89
LEU86	-0.81	1.16	-1.72	0.43	0.22
TRP87	-1.71	1.77	-2.25	0.23	0.14
GLY88	1.61	-0.58	1.09	0.97	0.95
ARG89	0.25	-0.27	0.17	0.67	0.65
CYS90	-0.16	0.39	-1.18	0.58	0.32
VAL91	0.28	-0.33	0.03	0.67	0.61
ALA92	0.12	-0.50	0.58	0.64	0.78
VAL93	-0.57	-0.15	-0.30	0.48	0.52
VAL94	-0.91	0.21	-0.78	0.41	0.41
VAL95	-0.44	-0.14	0.31	0.51	0.69
MET96	-0.79	-0.07	-0.25	0.44	0.53
VAL97	0.61	-0.22	-0.09	0.75	0.58
ALA98	-0.71	-0.06	-0.07	0.45	0.58
GLY99	1.09	-0.58	0.92	0.85	0.89
ILE100	1.14	-0.48	0.50	0.87	0.75
THR101	-0.56	-0.42	0.12	0.49	0.64
SER102	-0.98	-0.26	-0.13	0.39	0.57
PHE103	-0.2	-0.60	0.24	0.57	0.67
GLY104	1.29	-0.61	0.87	0.90	0.87
LEU105	0.95	-0.46	0.28	0.82	0.68
VAL106	0.94	-0.42	0.56	0.82	0.77
THR107	0.06	-0.52	0.56	0.63	0.77
ALA108	-0.76	-0.32	0.39	0.44	0.72
ALA109	0.29	-0.44	0.38	0.68	0.71
LEU110	0.82	-0.58	0.96	0.79	0.90
ALA111	-1.01	-0.41	0.01	0.39	0.60
THR112	0	-0.47	0.56	0.61	0.77
TRP113	-0.33	-0.28	0.42	0.54	0.73
PHE114	1.13	-0.55	0.73	0.86	0.82
VAL115	-0.9	-0.40	0.12	0.41	0.64
GLY116	-1.43	0.29	-0.94	0.29	0.37
ARG117	-1.1	-0.49	-0.28	0.37	0.53
GLU118	-0.06	-0.51	0.48	0.60	0.74
GLN119	-0.76	-0.66	-0.13	0.44	0.57

* Conservazione assegnata da ConSurf sulla base dell'allineamento ottenuto da ConSurf.
** Conservazione assegnata da ConSurf sulla base dell'allineamento ottenuto da CAMPO.

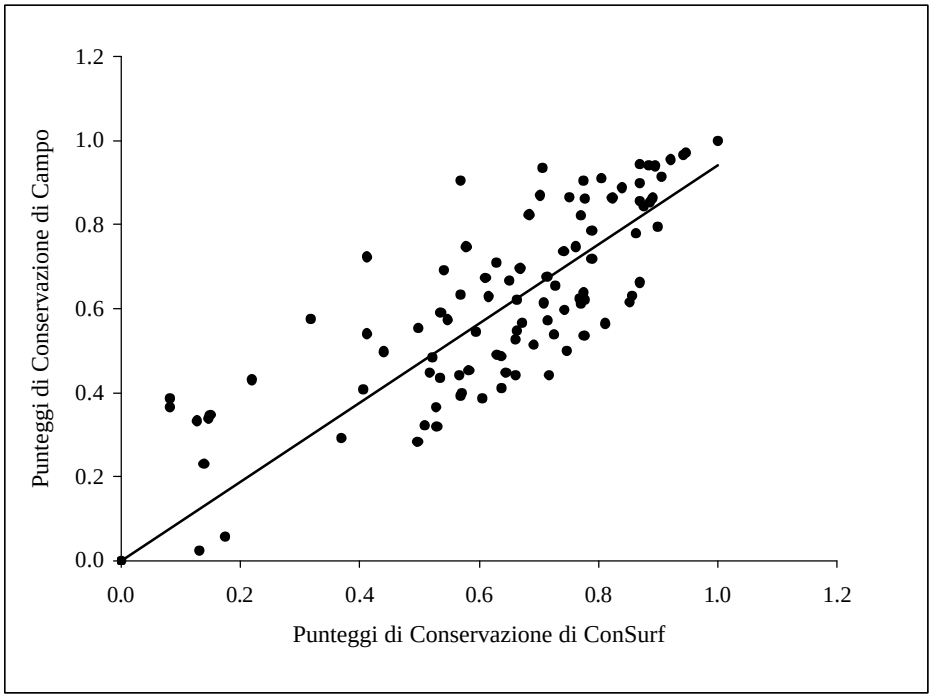


Fig. 5.12. Correlazione tra i punteggi di conservazione evolutiva ottenuti da CAMPO e da ConSurf per 1BL8.

E' interessante notare che ConSurf ha individuato, rispetto a CAMPO, una ulteriore regione C-terminale della struttura (Val 115, Arg 117, Glu 118 and Gln 119) alla quale è stato assegnato un alto punteggio di conservazione evolutiva (Tab. 5.IV), nonostante la presenza di 27 *gap* nelle 39 posizioni dell'allineamento multiplo. Evidentemente, in questo caso, ConSurf ha sottostimato la presenza di inserzioni e delezioni nell'allineamento multiplo di sequenza, con la conseguente errata identificazione di una regione evolutivamente conservata C-terminale. Inserzioni e delezioni sono infatti spesso indicative di regioni non conservate della proteina. Con l'utilizzo dello stesso allineamento multiplo di sequenza, le uniche differenze tra i due

programmi sembrano quindi consistere nel trattamento di inserzioni e delezioni.

SCR_FIND (http://schubert.bio.uniroma1.it/SCR_FIND) e CHC_FIND (http://schubert.bio.uniroma1.it/CHC_FIND) rappresentano le implementazioni degli algoritmi utilizzati per l'identificazione delle regioni strutturalmente conservate (SCRs) e dei contatti idrofobici conservati (CHCs) in famiglie e superfamiglie proteiche per le quali almeno due strutture tridimensionali sono state determinate sperimentalmente (Paragrafi 4.2 e 4.3).

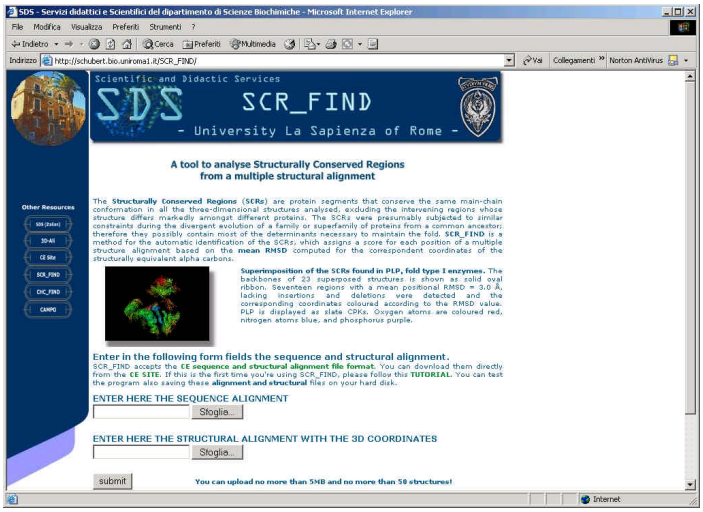


Fig. 5.13. La pagina iniziale di SCR_FIND

Fig. 5.14. La pagina iniziale di CHC_FIND

Sebbene oggi esistano molti programmi in rete capaci di analizzare relazioni strutturali (Schindyalov & Bourne, 1998; Sobolev *et al.*, 1999), lo sviluppo di un *server* dedicato all'estrazione delle SCRs e dei CHCs a partire da strutture proteiche allineate non è ancora disponibile (Fig. 5.13 e Fig. 5.14).

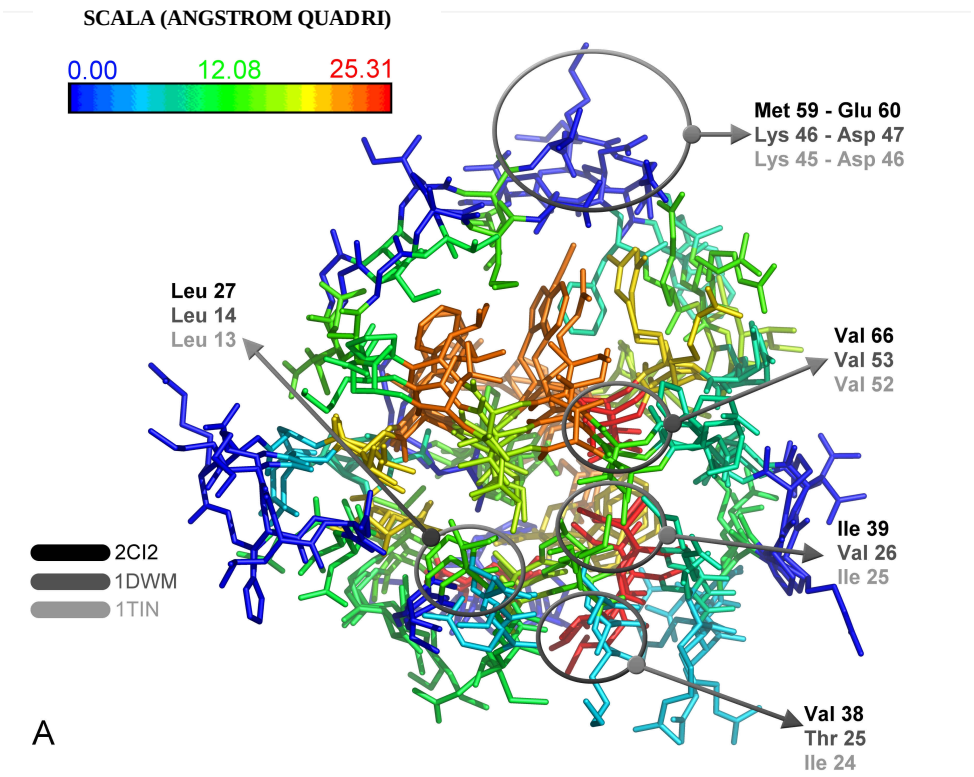
L'efficacia dei due programmi sarà di seguito valutata analizzando due famiglie il cui processo di ripiegamento e la cui stabilità è stata ampiamente caratterizzata: gli inibitori della tripsina e le proteine leganti il coenzima A.

Il processo di ripiegamento di queste piccole proteine globulari è un evento spontaneo *in vitro*, che ha luogo con un meccanismo di reazione a due stati (Kragelund *et al.*, 1999). Queste reazioni sono caratterizzate dalla presenza di un singolo stato di transizione che separa lo stato ripiegato da quello non ripiegato della proteina, con l'assenza di stati intermedi (Ptitsyn, 1991). E' stato suggerito che le interazioni nello stato di transizione sono simili a quelle dello stato nativo, con i residui coinvolti che formano un centro di nucleazione conservato (Shakhnovich *et al.*, 1996). Fino ad oggi, sono stati utilizzati prevalentemente approcci di mutagenesi sito-specifica per analizzare il meccanismo di ripiegamento di queste e molte altre piccole proteine globulari (Milla *et al.*, 1997; Jackson *et al.*, 1993; Lopez-Hernandez & Serrano, 1996).

L'inibitore 2 della chimotripsina (CI2; codice PDB: 2CI2; McPhalen & James, 1987), l'inibitore della tripsina da *Linum usitatissimum* (codice PDB: 1DWM; Cierpicki & Otlewski, 2000), e l'inibitore della tripsina da *Cucurbita maxima* (codice PDB: 1TIN; Krishnamoorthi *et al.*, 1990) sono proteine con un singolo dominio che condividono lo stesso ripiegamento. Grazie a SCR_FIND e CHC_FIND è stato possibile identificare un nucleo di interazioni conservate nelle SCRs che comprendono l'α-elica N-terminale ed

il foglietto β C-terminale di queste proteine (Fig. 5.15 A). Le interazioni idrofobiche maggiormente conservate sono costituite da (la numerazione si riferisce alla struttura 2CI2): leucina 27 con valina 38 (23.9 $Å^2$), ed isoleucina 39 con valina 66 (25.3 $Å^2$). Altri residui idrofobici mostrano schemi conservati di interazione: triptofano 24 ed alanina 35 con leucina 27 (19.8 $Å^2$ e 17.0 $Å^2$, rispettivamente), valina 50 con leucina 68 (17.7 $Å^2$), valina 70 con isoleucina 76 (17.7 $Å^2$), leucina 68 con isoleucina 76 (17.0 $Å^2$), e leucina 51 con fenilalanina 69 (22.7 $Å^2$). Molti di questi contatti sono stati precedentemente identificati come intermedi del ripiegamento, utilizzando approcci di ingegneria proteica. E' stato dimostrato che la complementazione di frammenti peptidici per originare una struttura simile a quella nativa avviene solo quando il taglio della proteina si trova nell'ansa di legame della proteasi, tra le posizioni metionina 59 e glutammato 60. In accordo con questa osservazione, questa regione non è coinvolta in alcun contatto idrofobico.

A differenza degli inibitori della tripsina, le proteine che legano il coenzima A (ACBPs) sono strutture costituite prevalentemente da α-eliche. Una ricerca iniziale con il programma in rete CE (Shindyalov & Bourne, 1998), ha permesso di identificare cinque strutture con il medesimo ripiegamento (Tab. 5.V). SCR_FIND ha estratto dalle strutture tridimensionali sovrapposte quattro SCRs, corrispondenti alle quattro α-eliche che costituiscono il nucleo proteico (Fig. 5.15 B).


SCALA (ANGSTROM QUADRI)
0.00
12.08
25.31
Met 59 - Glu 60
Lys 46 - Asp 47
Lys 45 - Asp 46
Leu 27
Leu 14
Leu 13
Val 66
Val 53
Val 52
Ile 39
Val 26
Ile 25
Val 38
Thr 25
Ile 24
2CI2
1DWM
1TIN
A

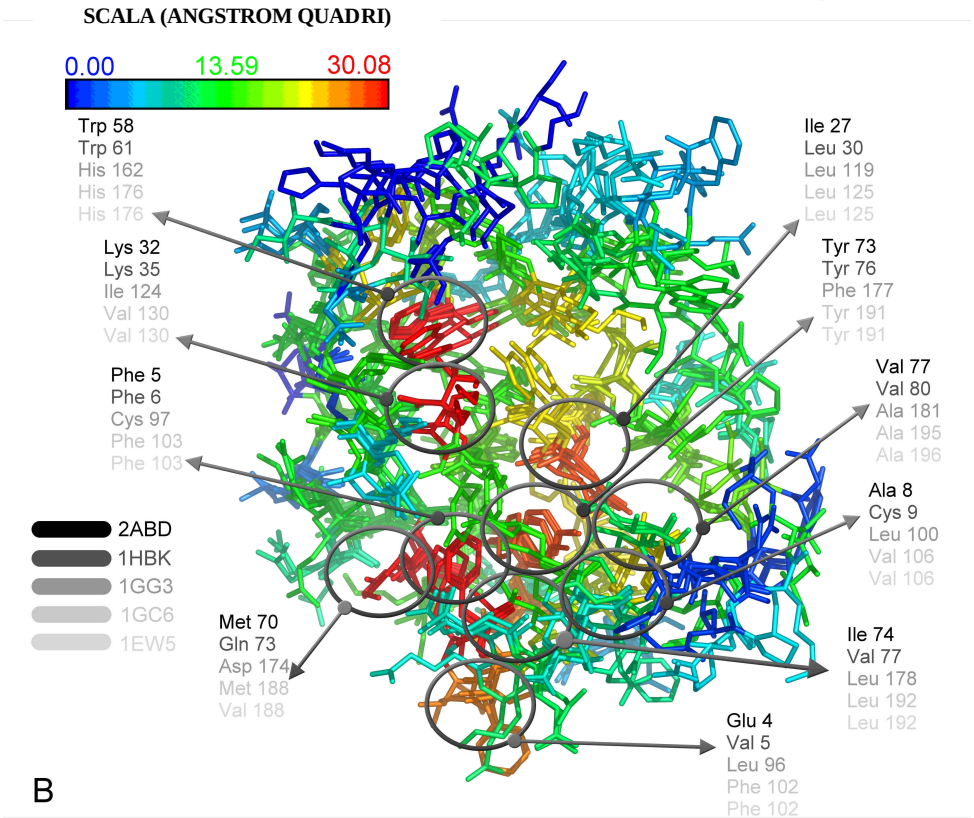

SCALA (ANGSTROM QUADRI)
0.00
13.59
30.08
Trp 58
Trp 61
His 162
His 176
His 176
Lys 32
Lys 35
Ile 124
Val 130
Val 130
Phe 5
Phe 6
Cys 97
Phe 103
Phe 103
Met 70
Gln 73
Asp 174
Met 188
Val 188
Ile 27
Leu 30
Leu 119
Leu 125
Leu 125
Tyr 73
Tyr 76
Phe 177
Tyr 191
Tyr 191
Val 77
Val 80
Ala 181
Ala 195
Ala 196
Ala 8
Cys 9
Leu 100
Val 106
Val 106
Ile 74
Val 77
Leu 178
Leu 192
Leu 192
Glu 4
Val 5
Leu 96
Phe 102
Phe 102
2ABD
1HBK
1GG3
1GC6
1EW5
B

Fig. 5.15 (pagina precedente). Esempio di risultati ottenuti con SCR_FIND e CHC_FIND. (A) Inibitori della tripsina e (B) proteine leganti il coenzima A. Le strutture tridimensionali sono colorate in accordo al valore medio di superficie di contatto apolare. I residui coinvolti nei contatti più intensi sono evidenziati.

Tabella 5.V. Confronto di RMSD e % di identità tra ACBPs.

Molecola (PDB)	**2ABD**	**1HBK**	**1GG3**	**1GC6**
1HBK	2.1 28.6			
1GG3	2.7 7.5	2.7 16.0		
1GC6	2.9 22.0	3.1 23.4	1.6 31.2	
1E5W	3.0 19.8	3.1 23.4	2.0 30.1	2.2 67.0

*RMSD e % di identità sono riportati sulla sinistra e sulla destra di ogni cella, rispettivamente.

A partire da questa impalcatura comune, CHC_FIND ha identificato un insieme di interazioni formate dalle eliche N- e C- terminali del nucleo strutturale; le interazioni coinvolgono i residui (la numerazione si riferisce a 2ABD): glutammato 4 con isoleucina 74 (24.9 $Å^2$), fenilalanina 5 con metionina 70 e tirosina 73 (30.0 $Å^2$ e 27.7 $Å^2$), alanina 8 con valina 77 (21.3 $Å^2$). Queste interazioni formano cerniere tra la prima e l'ultima elica del nucleo proteico. Studi di mutagenesi sito-specifica hanno identificato i

medesimi residui (con l'eccezione della metionina 70, che non è stata mutata) come principalmente responsabili della formazione di strutture simili alla nativa, nel processo di ripiegamento delle ACBPs (Fersht, 1997).

DISCUSSIONE

6.1 Evoluzione molecolare degli enzimi PLP-dipendenti di tipo I

Il lavoro presentato è stato principalmente indirizzato all'identificazione delle caratteristiche strutturali rimaste invariate per lunghi periodi evolutivi negli enzimi PLP-dipendenti di tipo I. Questa superfamiglia è particolarmente adatta a questo tipo di analisi, dal momento che i suoi rappresentanti sono legati da una lunga storia di evoluzione divergente. E' stato proposto, infatti, che questi enzimi fossero probabilmente già presenti nella cellula universale ancestrale circa 1500 milioni di anni fa (Mehta & Christen, 1998). Questa lunga storia evolutiva ha fatto sì che, nonostante l'omologia strutturale tra i diversi membri di questa superfamiglia sia ancora riconoscibile, il grado di similarità di sequenza non sia sufficiente per stabilire una origine comune. Per tali ragioni, questa superfamiglia può essere considerata *per se* un modello di plasticità evolutiva proteica.

La somiglianza strutturale nel campione raccolto è distribuita secondo le differenti funzioni locali svolte dai motivi di struttura secondaria. Circa il 30% dei residui formano un nucleo comune di elementi di struttura secondaria ben conservato. La conservazione strutturale osservata è probabilmente dovuta alle restrizioni spaziali imposte dalle comuni modalità di legame del PLP, che viene a trovarsi all'interno di una tasca idrofobica all'interfaccia tra le due subunità. Fatta eccezione per questo nucleo strutturale comune, la lunghezza degli elementi di struttura secondaria regolare e delle anse differisce in maniera sostanziale tra membri evolutivamente distanti; tale considerazione risulta evidente anche

dall'osservazione dell'allineamento multiplo di 921 sequenze, dove sono molto rari i blocchi conservati privi di inserzioni e delezioni. Probabilmente, in risposta alle modifiche dell'apparato catalitico, richieste per modulare la specificità di reazione e di substrato, grandi adattamenti strutturali hanno avuto origine durante l'evoluzione divergente di questa superfamiglia da un progenitore comune. Quasi certamente, le anse che circondano l'ingresso al sito attivo sono state principalmente coinvolte in queste modifiche strutturali (Contestabile *et al.,* 2001).

6.2 Contatti idrofobici conservati

Questo lavoro è stato indirizzato principalmente all'analisi della conservazione dei contatti idrofobici, all'interno delle regioni strutturalmente conservate che sono risultate dal confronto strutturale del campione utilizzato. La conservazione dei contatti idrofobici è il risultato della pressione selettiva esercitata durante l'evoluzione molecolare, per mantenere un ripiegamento funzionalmente attivo. L'analisi ha permesso di individuare tre insiemi di contatti idrofobici conservati: il primo ed il secondo gruppo di CHCs (Fig. 5.7) sono localizzati in prossimità di residui indispensabili al corretto posizionamento del cofattore e dei substrati all'interno del sito attivo; il terzo gruppo forma una cerniera tra le SCR α_1 ed α_{12}, che delimitano il dominio maggiore.

Per quanto riguardo il primo gruppo di CHCs, è considerevole la separazione tra i residui coinvolti in un ruolo funzionale (interazione con il cofattore e modulazione della sua attività), tutti localizzati all'estremità di un nucleo strutturale comune (costituito dalle SCRs α_3, β_6, β_9, β_{10}, e β_{11}), ed i residui coinvolti in un ruolo strutturale (mantenimento della stabilità

strutturale attraverso la formazione di CHCs), posizionati invece alla base della stessa unità funzionale (Fig. 6.1). Questa disposizione spaziale e funzionale, che comprende un'impalcatura stabile comune che si ripiega attorno ad un nucleo funzionalmente variabile di residui, può essere osservata in molte altre unità strutturali "di successo", da un punto di vista evolutivo, ovvero largamente diffuse in natura (per esempio, i domini *tim-barrell* o *Ig-like*; Nagano *et al.*, 2002; Selvaraj & Gromiha, 2003) e sembra fornire una soluzione al compromesso tra stabilità della struttura e plasticità della funzione, aumentando la specificità di reazione e di substrato, senza compromettere il ripiegamento e la conformazione strutturali (Wierenga, 2001; Todd *et al.*, 2001; Nagano *et al.*, 2002).

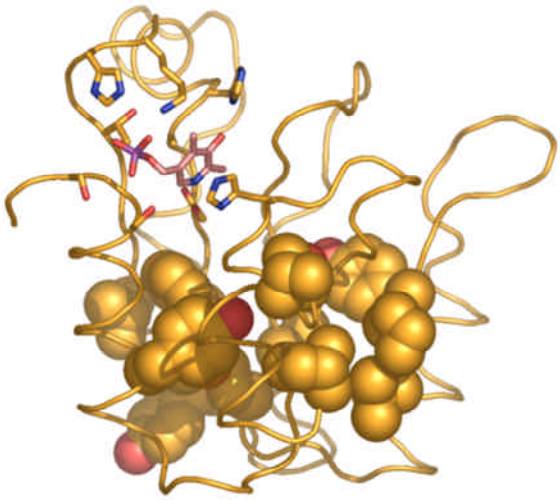


Fig. 6.1. Nucleo strutturale comune ai membri analizzati della superfamiglia degli enzimi PLP-dipendenti di tipo I, costituito dalle SCRs α_3, β_6, β_9, β_{10}, e β_{11}. Sono state evidenziate le catene laterali dei residui interagenti con il PLP e, utilizzando una rappresentazione "a spazio pieno", le catene laterali dei residui implicati nella formazione dei contatti idrofobici conservati. Il modello strutturale di riferimento utilizzato per la rappresentazione è la serina idrossimetiltrasferasi da *H. sapiens* (PDB 1BJ4).

In contrasto apparente con i due insiemi precedentemente descritti, il gruppo di CHCs localizzato nelle eliche α_1 e α_{12} non sembra essere coinvolto nel corretto posizionamento o nella stabilità di alcun residuo del sito attivo. L'esame della rete di contatti ha messo in evidenza che i CHCs giacciono lungo ciascun lato dell'elica, formando una cerniera nascosta al solvente nelle posizioni *i*, *i* + 4 ed *i* + 7. Questo particolare schema, assolutamente conservato, di contatti tra residui è stato precedentemente identificato da Hill *et al.* (2002) e da Ptitsyn (1998) nell'analisi compiuta sulla superfamiglia delle citochine e sui citocromi di tipo *c*, rispettivamente. In entrambi gli studi è stato concluso che questi residui erano di importanza critica nel ripiegamento proteico.

Nel caso degli enzimi PLP-dipendenti, precedenti studi sperimentali hanno suggerito la presenza di tre nuclei strutturali responsabili del corretto schema di ripiegamento e della stabilità degli enzimi di tipo I. Herold *et al.* (1991) hanno dimostrato che il dominio maggiore, legante il PLP, dell'aspartato amminotrasferasi da *E. coli*, corrispondente al primo ed al terzo dominio nei quali sono stati individuati i CHCs, è capace di ripiegarsi in maniera autonoma sia *in vivo*, sia *in vitro*, e di legare il PLP. Più recentemente, Fu *et al.* (2003) hanno proposto che il meccanismo di ripiegamento della serina idrossimetiltrasferasi da *E. coli* può essere suddiviso in due passaggi: una prima fase veloce nella quale i due domini, che corrispondono al primo ed al secondo dominio nei quali sono presenti i CHCs, si ripiegano nella loro conformazione nativa; una lenta fase finale nella quale un segmento interdominio della catena polipeptidica, che comprende l'elica α_{12}, si ripiega nella sua conformazione nativa, interagendo con l'elica α_1 N-terminale del dominio maggiore. Si pensa che questo ultimo

passaggio sia implicato nel legame del PLP (Fu *et al.* , 2003). L'analisi condotta avvalora questa ipotesi e suggerisce una possibile spiegazione meccanicistica a questi studi sperimentali, servendo da base per ulteriori esperimenti tesi a stabilire la correlazione struttura-funzione, ed a comprendere il ruolo dei singoli residui e delle interazioni a coppie nel ripiegamento e nella stabilità di questa superfamiglia proteica.

6.3 Relazione tra conservazione di sequenza e conservazione strutturale

Uno degli obiettivi principali di questo studio è stato quello di determinare se le restrizioni strutturali comuni, necessarie all'avvicinamento dei residui interagenti all'interno del nucleo strutturale degli enzimi PLP-dipendenti di tipo I, sia riflesso in qualche modo da uno schema di conservazione di sequenza, osservato nelle posizioni dell'allineamento multiplo di questa superfamiglia. Il grafico del grado di conservazione evolutiva media di due residui interagenti delle SCRs in funzione del grado di contatto idrofobico medio della loro frazione apolare può essere descritto da una relazione lineare (r = 0.70).

Nell'analisi condotta, è stata considerata la media della conservazione evolutiva di coppie di residui interagenti, in luogo della conservazione della singola posizione. Un vantaggio considerevole nel considerare la conservazione a coppie è che questa permette di prendere in considerazione co-mutazioni che possono avvenire nella sequenza amminoacidica durante l'evoluzione. E' importante sottolineare, difatti, che le posizioni conservate non sono invarianti; al contrario, nel confronto di differenti strutture possono essere osservate mutazioni correlate. Di conseguenza, sembra che ciò che è realmente "conservato" sia la posizione spaziale dell'interazione ed il suo

effetto idrofobico, piuttosto che l'identità specifica delle catene laterali che partecipano alla formazione di un CHC.

Sebbene i 23 enzimi PLP-dipendenti a struttura nota presi in considerazione siano evolutivamente molto distanti, è possibile comunque individuare un profilo strutturale di contatti idrofobici conservati, l'importanza dei quali, nella stabilizzazione del ripiegamento nativo, è sostenuta da una conservazione evolutiva preferenziale nel campione di sequenze considerato.

6.4 Presenza del nucleo strutturale conservato in altre superfamiglie proteiche

Le regioni strutturalmente conservate estratte dal campione di enzimi PLP-dipendenti possono essere considerate come un'eredità strutturale della macromolecola ancestrale dalla quale questi enzimi si sono evoluti. In particolare, cinque delle diciassette SCRs trovate (α_3, β_6, β_9, β_{10}, e β_{11}) sono principalmente implicate nella costituzione del nucleo proteico che circonda il PLP, e dal quale provengono la maggior parte dei residui indispensabili al legame del cofattore. E' plausibile assumere che questo nucleo strutturale rappresenti il lascito strutturale più antico, da un punto di vista evolutivo, dell'enzima ancestrale.

Recentemente, Christen & Mehta (2001) hanno proposto che i cofattori organici apparvero sulla scena evolutiva prima della formazione dei rispettivi apoenzimi, e che un metabolismo primitivo controllato cineticamente dai cofattori fornì i primi "blocchi strutturali" necessari, a loro volta, ad assistere in maniera sempre più efficiente i cofattori nelle loro funzioni catalitiche. In questo mondo prebiotico, gli enzimi ancestrali avrebbero quindi

rappresentato primitive impalcature proteiche di cofattori organici. Una volta formatisi, questi apoenzimi ancestrali avrebbero potuto dare origine a nuovi oloenzimi, interagendo con i preesistenti cofattori disponibili.

A partire da questa considerazione di base, si è tentato di individuare la possibile presenza di questa unità strutturale in proteine apparentemente non correlate alla superfamiglia analizzata. Tale ricerca, eseguita attraverso l'utilizzo del programma CE, ha permesso di identificare alcune proteine contenenti questo motivo strutturale, oltre agli enzimi PLP dipendenti (Tab. 6.I).

In molti casi, sembra che questi elementi strutturali siano apparsi nel corso dell'evoluzione in differenti occasioni: gli elementi di struttura secondaria che costituiscono il nucleo proteico non sono ordinati in maniera sequenziale, provenendo da regioni differenti della catena polipeptidica od avendo direzioni opposte della catena; inoltre, sono spesso presenti estese inserzioni e delezioni. Ciò nonostante, in alcuni casi (ad esempio, il dominio isomerasico della glucosamina-6-fosfato sintasi, PDB 1MOR; Fig. 6.2), la colinearità degli elementi di struttura secondaria solleva dei dubbi sulle relazioni evolutive che sussistono tra queste unità strutturali. Inoltre, il nucleo proteico trovato negli enzimi PLP-dipendenti costituisce, in altre proteine, (la proteina legante il D-ribosio, PDB 1URP, o CheY, PDB 1UDR) un dominio indipendente capace di ripiegarsi in maniera autonoma.

Tabella 6.I. Lista degli enzimi trovati in banca dati PDB che presentano il dominio in comune con gli enzimi PLP dipendenti di ripiegamento di tipo I.

Codice PDB	Nome enzima	Cofattore	Funzione
3ENL	enolasi	-	Metabolismo cellulare
1DBQ	Proteina regolatrice legante il DNA	-	Regolatore trascrizionale
1ID1	Dominio citoplasmatico di un canale del potassio	-	Canale transmembrana per il potassio
1MOR	Dominio isomerasico della glucosamina 6 fosfato sintasi	-	Isomerizza l'α D glucosio 6 fosfato in D fruttosio 6 fosfato
1POW	piruvato ossidasi	FAD e tiamina pirofosfato	Metabolismo cellulare, decarbossila il piruvato
1QO0	Complesso del recettore amidico dell'operone dell'amidasi.	-	Regola negativamente l'espressione dll'operone dell'amidasi.
1SRR	Fosfotransferasi implicata nella sporulazione	-	
1TLF	Repressore Lac	-	Proteina legante il DNA
1UDR	Che y	-	Proteina traduttrice del segnale che agisce come sensore chemiotattico.
1URP	Proteina legante il D-ribosio	-	Proteina coinvolta nel trasporto dello D-ribosio.
2GBP	Proteina legante il D-glucosio e il D-galattosio	-	Agisce come trasduttore del segnale chemiotattico
1D4O	Fosfoidrogenasi NADPH dipendente	FAD	Converte il NADPH prodotto dalle reazioni cellulari in NADH.
1ETU	Ef TU	GDP	Fattore trascrizionale di elongazione, che lega la guanosina 5 difosfato ed il DNA
1DC7	Proteina regolatrice dell'azoto	-	"Bottone" per la traduzione del segnale nel metabolismo dell'azoto

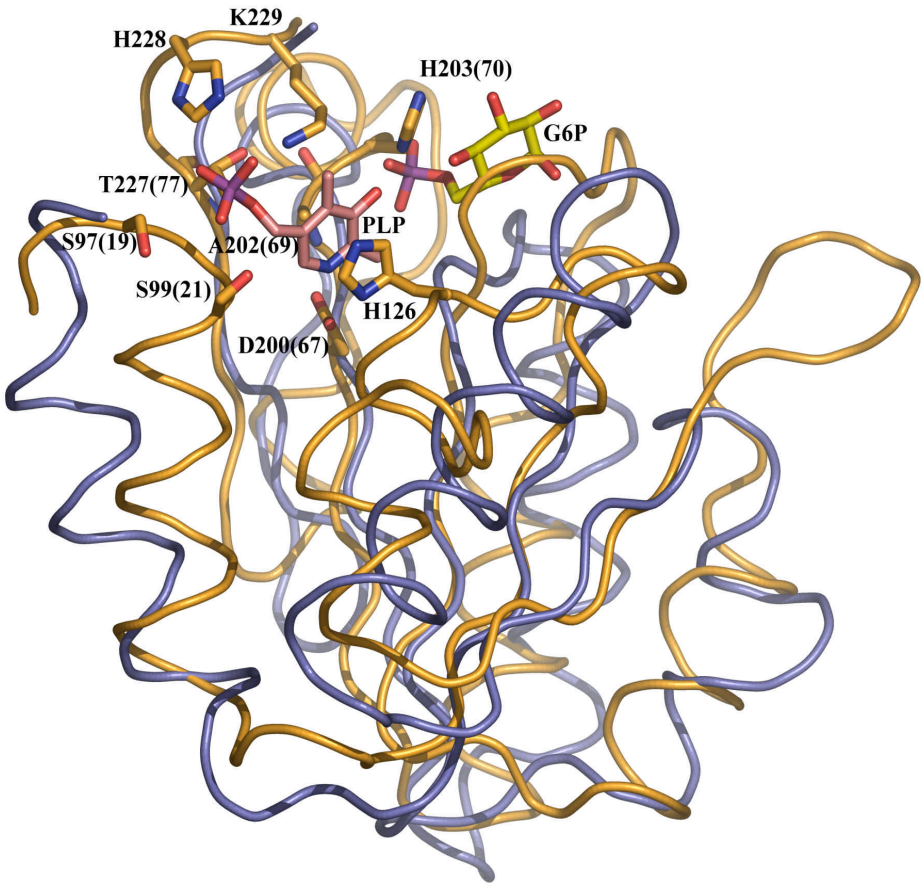


Fig. 6.2. Sovrapposizione del nucleo strutturale comune degli enzimi di tipo I con il dominio isomerasico della glucosammina-6-fosfato sintasi (1MOR). La serina idrossimetiltrasferasi da *E. coli* (eSHMT) è stata utilizzata per rappresentare il nucleo proteico comune (arancione). 1MOR è rappresentata in azzurro. I residui implicati nel legame del PLP sono mostrati come bastoncelli, ed indicati in base alla numerazione della eSHMT. I numeri tra parentesi indicano le posizioni assolute della Fig. 5.2. Il PLP ed il glucosio-D-6-fosfato (G6P) sono mostrati in rosa e giallo, rispettivamente. Gli atomi di ossigeno sono colorati in rosso, l'azoto in blu ed il fosforo viola.

CONCLUSIONI

Il lavoro presentato nasce dal tentativo di contribuire, almeno in parte, a dare una risposta agli interrogativi che ci siamo posti inizialmente, sulla possibilità di riuscire ad estrarre dai dati di sequenza e strutturali disponibili relativamente ad una superfamiglia proteica l'informazione necessaria a determinare il ripiegamento e mantenerne la stabilità dei suoi membri. In particolare, focalizzando l'analisi sulla superfamiglia degli enzimi PLP-dipendenti di tipo I, ci siamo chiesti se era possibile individuare, in proteine così evolutivamente distanti, un profilo idrofobico simile, conservato al livello di struttura primaria e terziaria.

La risposta sembra affermativa: esiste un assetto di contatti idrofobici conservato, a dispetto della bassissima conservazione di sequenza osservata nei membri di questa superfamiglia. Inoltre, esiste una correlazione statisticamente significativa tra la conservazione evolutiva e la superficie di contatto apolare media dei residui coinvolti nella formazione di CHCs, al punto che alcuni di essi mostrano una conservazione confrontabile con quella dei residui che interagiscono con il cofattore.

L'importanza della relazione tra residui evolutivamente conservati e contatti idrofobici è stata inizialmente sottolineata da Shakhnovich *et al.*, (1996). Questi autori hanno ipotizzato che il "nucleo di ripiegamento" (*folding nucleus*), un sottoinsieme della struttura nativa nel quale i residui stabiliscono contatti idrofobici determinanti per il corretto ripiegamento della proteina, può essere descritto come un insieme di residui evolutivamente conservati e privi, apparentemente, di un ruolo funzionale. Alla luce di questa teoria, possiamo ipotizzare che i tre gruppi identificati di contatti idrofobici

conservati rappresentino i “nuclei di ripiegamento” della superfamiglia di enzimi PLP-dipendenti di tipo I.

Comunque, il dibattito sulla relazione tra conservazione di sequenza e ripiegamento proteico è ancora aperto, e saranno necessari certamente ulteriori studi sperimentali per comprendere il coinvolgimento dei CHCs nel ripiegamento, in questa ed in altre superfamiglie. In questo senso riteniamo, infine, che le strategie e gli algoritmi sviluppati e resi disponibili alla comunità scientifica, sotto forma di programmi accessibili dalla rete, potranno essere d’ aiuto nella progettazione e nell’interpretazione di esperimenti di ripiegamento proteico.

BIBLIOGRAFIA

Alexeev, D., Alexeeva, M., Baxter, R.L., Campopiano, D.J., Webster, S.P., and Sawyer, L. 1998. The crystal structure of 8-amino-7-oxononanoate synthase: a bacterial PLP-dependent, acyl-CoA-condensing enzyme. *J. Mol. Biol.* **284:** 401-419.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-3402.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223-230.

Armon, A., Graur, D., and Ben-Tal, N. 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307:** 447-463.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Research*. **28:** 235-242.

Blaber, M., Zhang, X.J., and Matthews, B.W. 1993. Structural basis of amino acid alpha helix propensity. *Science* **260:** 1637-40.

Britton, K.L., Baker, P.J., and Borges, K.M.M. 1995. Insights into thermal stability from a comparison of the glutamate dehydrogenases from *Pyrococcus furiosus* and *Thermococcus litoralis. Eur. J. Biochem.* **229:** 688-695.

Burkhard, P., Dominici, P., Borri-Voltattorni, C., Jansonius, J.N., and Malashkevich, V.N. 2001. Structural insight into Parkinson's disease

treatment gained from drug-inhibited dopa decarboxylase. *Nat. Struct. Biol.* **8:** 963-967.

Chen, J., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler-Bauer, A., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Rao, B.S., Panchenko, A.R., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J., and Bryant, S.H. 2003. MMDB: Entrez's 3D-structure database. *Nucleic Acids Research.* **31:** 474-477.

Chothia, C., and Lesk, A. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823-826.

Christen, P., and Mehta, P.K. 2001. From cofactor to enzymes. The molecular evolution of pyridoxal-5'-phosphate-dependent enzymes. *Chem. Rec*. **1:** 436-447.

Cierpicki, T., and Otlewski, J. 2000. Determination of a high precision structure of a novel protein (*Linum usitatissimum* trypsin inhibitor LUTI) using computer-aided assignment of NOESY cross-peaks. *J. Mol. Biol.* **302:** 1179-1192.

Clausen, T., Huber, R., Laber, B., Pohlenz, H.D., and Messerschmidt, A. 1996. Crystal structure of the pyridoxal-5'-phosphate dependent cystathionine beta-lyase from *Escherichia coli* at 1.83 Å. *J. Mol. Biol.* **262:** 202-224.

Clausen, T., Kaiser, J.T., Steegborn, C., Huber, R., and Kessler, D. 2000. Crystal structure of the cystine C-S lyase from *Synechocystis*: stabilization of cysteine persulfide for FeS cluster biosynthesis. *Proc. Natl. Acad. Sci. USA.* **97:** 3856-3861.

Clausen, T., Schlegel, A., Peist, R., Schneider, E., Steegborn, C., Chang, Y.S., Haase, A., Bourenkov, G.P., Bartunik, H.D., and Boos, W. 2000. X-ray structure of MalY from *Escherichia coli*: a pyridoxal 5'-phosphate-dependent enzyme acting as a modulator in mal gene expression. *EMBO J.* **19:** 831-842.

Contestabile, R., Paiardini, A., Pascarella, S., di Salvo, M.L., D'Aguanno, S., and Bossa, F. 2001. l-Threonine aldolase, serine hydroxymethyltransferase and fungal alanine racemase. A subgroup of strictly related enzymes specialized for different functions. *Eur. J. Biochem.* **268:** 6508-6525.

Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T., and MacKinnon, R. 1998. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science* **280:** 69-77.

Drabløs, F. 1999. Clustering of non-polar contacts in proteins. *Bioinformatics.* **15:** 501-509.

Dunathan, H.C. 1966. Conformation reaction specificity in pyridoxal-phosphate enzymes. *Proc. Natl. Acad. Sci.* USA. **55:** 712-716.

Eads, J.C., Beeby, M., Scapin, G., Yu, T.W., and Floss, H.G. 1997. The crystal structure of 3-amino-5-hydroxybenzoic acid Ahba synthase. *Biochemistry.* **38:** 9840-9849.

Fersht, A. 1997. Nucleation mechanism of protein folding. *Curr. Opin. Struct. Biol.* **7:** 10-14.

Fu, T.F., Boja, E.S., Safo, M.K., and Schirch, V. 2003. Role of proline residues in the folding of serine hydroxymethyltransferase. *J. Biol. Chem.* **278:** 31088-31094.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., Ben-Tal, N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19:** 163-164.

Grishin, N.V., Phillips, M.A., and Goldsmith, E.J. 1995. Modeling of the spatial structure of eukaryotic ornithine decarboxylases. *Protein Sci.* **4:** 1291-1304.

Gromiha, M.M., Pujadas, G., Magyar, C., Selvaraj, S., and Simon, I. 2004. Locating the stabilizing residues in $(\alpha/\beta)_8$ barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. *Proteins* **55:** 316-329.

Gunasekaran, K., Hagler, A.T., and Gierasch, L.M. 2004. Sequence and structural analysis of cellular retinoic acid-binding proteins reveals a network of conserved hydrophobic interactions. *Proteins* **54:** 179-194.

Harrison, P. Percolation Theory. 2001. *In*: Computational Methods in Physics Chemistry and Biology: An Introduction. New York: John Wiley & Sons Inc; D L Purich, editor. p. 209-224.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* **89:** 10915-10919.

Hennig, M., Grimm, B., Contestabile, R., John, R.A., and Jansonius, J.N. 1997. Crystal structure of glutamate-1-semialdehyde aminomutase: an alpha2-dimeric vitamin B6-dependent enzyme with asymmetry in structure and active site reactivity. *Proc. Natl. Acad. Sci. USA.* **94:** 4866-4871.

Herold, M., Leistler, B., Hage, A., Luger, K., and Kirschner K. 1991. Autonomous folding and coenzyme binding of the excised pyridoxal

5'-phosphate binding domain of aspartate aminotransferase from *Escherichia coli*. *Biochemistry* **30:** 3612-20.

Hester, G., Stark, W., Moser, M., Kallen, J., Markovic-Housley, Z., and Jansonius, J.N. 1999. Crystal structure of phosphoserine aminotransferase from *Escherichia coli* at 2.3 A resolution: comparison of the unligated enzyme and a complex with alpha-methyl-l-glutamate. *J. Mol. Biol.* **286:** 829-850.

Hill, E.E., Morea, V., and Chothia, C. 2002. Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. *J. Mol. Biol.* **322:** 205-233.

Hohenester, E., Keller, J.W., and Jansonius, J.N. 1994. An alkali metal ion size-dependent switch in the active site structure of dialkylglycine decarboxylase. *Biochemistry* **33:** 13561-13570.

Holm, L., and Sander, C. 1998. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14:** 423-429.

Isupov, M.N., Antson, A.A., Dodson, E.J., Dodson, G.G., Dementieva, I.S., Zakomirdina, L.N., Wilson, K.S., Dauter, Z., Lebedev, A.A., and Harutyunyan, E.H. 1998. Crystal structure of tryptophanase. *J. Mol. Biol.* **276:** 603-623.

Jackson, S.E., elMasry, N., and Fersht, A.R. 1993. Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry* **32:** 11270-11278.

Jansonius, J. 1998. Structure, evolution and action of vitamin B_6-dependent enzymes. *Curr. Opin. Struct. Biol.* **8:** 759-769.

Jin, H., and Martin, C. 1999. Multifunctionality and diversity within the plant MYB-gene family. *Plant Mol Biol.* **41:** 577-85.

John, R.A. 1995. Pyridoxal phosphate-dependent enzymes. *Biochim. Biophys. Acta* **1248:** 81-96.

Junker, V., Contrino, S., Fleischmann, W., Hermjakob, H., Lang, F., Magrane, M., Martin, M.J., Mitaritonna, N., ÒDonovan, C., and Apweiler, R. 2000. The role SWISS-PROT and TrEMBL play in the genome research environment. *J. Biotechnol.* **78:** 221-234.

Käck, H., Sandmark, J., Gibson, K., Schneider, G., and Lindqvist, Y. 1999. Crystal structure of diaminopelargonic acid synthase: evolutionary relationships between pyridoxal-5'-phosphate-dependent enzymes. *J. Mol. Biol.* **291:** 857-876.

Kaiser, J.T., Clausen, T., Bourenkow, G.P., Bartunik, H.D., Steinbacher, S., and Huber, R. 2000. Crystal structure of a NifS-like protein from *Thermotoga maritima*: implications for iron sulphur cluster assembly. *J. Mol. Biol.* **297:** 451-464.

Karlin, S., and Brocchieri, L. 1996. Evolutionary conservation of RecA genes in relation to protein structure and function. *J. Bacteriol.* **178:** 1881-1894.

Kielkopf, C.L., and Burley, S.K. 2002. X-ray structures of threonine aldolase complexes: structural basis of substrate recognition. *Biochemistry* **41:** 11711-11720.

Kragelund, B.B., Osmark, P., Neergaard, T.B., Schiodt, J., Kristiansen, K., Knudsen, J., and Poulsen, F.M. 1999. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat. Struct. Biol.* **6:** 594-601.

Krishnamoorthi, R., Gong, Y.X., and Richardson, M. A new protein inhibitor of trypsin and activated Hageman factor from pumpkin (*Cucurbita maxima*) seeds. *FEBS Lett.* **273:** 163-167.

Krupka, H.I., Huber, R., Holt, S.C., and Clausen, T. 2000. Crystal structure of cystalysin from *Treponema denticola*: a pyridoxal 5'-phosphate-dependent protein acting as a haemolytic enzyme. *EMBO J.* **19:** 3168-3178.

Kuettner, E.B., Hilgenfeld, R., and Weiss, M.S. 2002. The active principle of garlic at atomic resolution. *J. Biol. Chem.* **277:** 46402-46407.

Lesk, A., and Chothia, C. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136:** 225–270.

Lopez-Hernandez, E., and Serrano, L. 1996. Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein CI-2. *Fold Des.* **1:** 43-45.

McPhalen, C.A., and James, M.N. 1987. Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry* **26:** 261-269.

Mehta, P.K., and Christen, P. 1998. The molecular evolution of Pyridoxal-5'-phosphate-dependent enzymes. *In* Advances in Enzymology and Related Areas of Molecular Biology: Mechanism of Enzyme Action, Part B. D. L. Purich, editor. John Wiley & Sons, Inc. p. 129-184.

Milla, M.E., Brown, B.M., Waldburger, C.D., and Sauer, R.T. 1997. P22 Arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry* **34:** 13914-13919.

Miyata, T., Miyazawa, S., and Yashunaga, T. 1979. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12:** 219-236.

Momany, C., Ernst, S., Ghosh, R., Chang, N.L., and Hackert, M.L. 1995. Crystallographic structure of a PLP-dependent ornithine decarboxylase from *Lactobacillus* 30a to 3.0 A resolution. *J. Mol. Biol.* **252:** 643-655.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536-540.

Nagano, N., Orengo, C.A., and Thornton, J.M. 2002. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321:** 741-765.

Nakai, T., Okada, K., Akutsu, S., Miyahara, I., Kawaguchi, S., Kato, R., Kuramitsu, S., and Hirotsu, K. 1999. Structure of *Thermus thermophilus* HB8 aspartate aminotransferase and its complex with maleate. *Biochemistry* **38:** 2413-2424.

Noland, B.W., Newman, J.M., Hendle, J., Badger, J., Christopher, J.A., Tresser, J., Buchanan, M.D., Wright, T.A., Rutter, M.E., Sanderson, W.E., Muller-Dieckmann, H.-J., Gajiwala, K.S., and Buchanan, S.G. 2002. Structural studies of *Salmonella typhimurium* Arnb Pmrh aminotransferase: a 4-amino-4-deoxy-L-arabinose liposaccharide modifying enzyme. *Structure* **10:** 1569-1580.

Okamoto, A., Nakai, Y., Hayashi, H., Hirotsu, K., and Kagamiyama, H. 1998. Crystal structures of *Paracoccus denitrificans* aromatic amino acid aminotransferase: a substrate recognition site constructed by rearrangement of hydrogen bond network. *J. Mol. Biol.* **280:** 443-461.

Orengo, C.A., Pearl, F.M., and Thornton, J.M. 2003. The CATH domain structure database. *Methods Biochem. Anal.* **44:** 249-271.

Pascarella, S., and Argos, P. 1992. A data bank merging related protein structures and sequences. *Protein Eng*. **5:** 121-137.

Ptitsyn, O.B. 1991. How does protein synthesis give rise to the 3D-structure? *FEBS Lett*. **285:** 176-181.

Ptitsyn, O.B. 1998. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* **278:** 655-666.

Renwick, S.B., Snell, K., and Baumann, U. 1998. The crystal structure of human cytosolic serine hydroxymethyltransferase: a target for cancer chemotherapy. *Structure* **6:** 1105-1116.

Richardson, J.S., and Richardson, D.C. 1989. Principles and patterns of protein conformation. *In* Prediction of Protein Structure and the Principles of protein conformation. G.D. Fasman, editor. Plenum Press. p. 1-99.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* **12:** 85-94.

Sanchez, R., Pieper, U., Mirkovic, N., De Bakker, P.I.W., Wittenstein, E., and Šali, A. 2002. MODBASE: a database of annotated comparative protein structure models. *Nucleic Acids Res.* **28:** 255-259.

Schneider, G., Käck, H., and Lindqvist, Y. 2000. The manifold of vitamin B6 dependent enzymes. *Structure* **8:** 1-6.

Selvaraj, S., and Gromiha, M.M. 2003. Role of hydrophobic clusters and long-range contact networks in the folding of alpha/beta8 barrel proteins. *Biophys. J.* **84:** 1919-1925.

Shakhnovich, E., Abkevich, V., and Ptitsyn, O. 1996. Conserved residues and the mechanism of protein folding. *Nature* **379:** 96-98.

Shindyalov, I.N., and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*. **11:** 739-747.

Sivaraman, J., Li, Y., Larocque, R., Schrag, J.D., Cygler, M., and Matte, A. 2001. Crystal structure of histidinol phosphate aminotransferase HisC from *Escherichia coli*, and its covalent complex with pyridoxal-5'-phosphate and l-histidinol phosphate. *J. Mol. Biol.* **311:** 761-776.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., and Edelman, M. 1999. Automated analysis of interatomic contacts in proteins. *Bioinformatics* **15:** 327-332.

Steegborn, C., Messerschmidt, A., Laber, B., Streber, W., Huber, R., and Clausen, T. 1999. The crystal structure of cystathionine gamma-synthase from *Nicotiana tabacum* reveals its substrate and reaction specificity. *J. Mol. Biol.* **290:** 983-996.

Storici, P., Capitani, G., De Biase, D., Moser, M., John, R.A., Jansonius, J.N., and Schirmer, T. 1999. Crystal structure of gaba-aminotransferase, a target for antiepileptic drug therapy. *Biochemistry* **38:** 8628-8634.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673-4680.

Todd, A.E, Orengo, C.A., and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307:** 1113-1143.

Valdar WS. 2002. Scoring residue conservation. *Proteins* **48:** 227-241.

Vogt, G., Etzold, T., and Argos, P. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J. Mol. Biol.* **249:** 816-831.

Wierenga, R.K. 2001. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492:** 193-198.