

Università degli Studi di Padova

Facoltà di Scienze MM. FF. NN.

Laurea Specialistica in Biologia Molecolare



Analisi *in silico* della Pendrina

Relatore: **Prof. Silvio Tosatto**

Dipartimento di Biologia

Correlatore: **Dr. Alessandra Murgia**

Dipartimento di Pediatria

Laureanda: **Emanuela Leonardi**

Anno Accademico 2007-2008

Indice

Abstract.....	3
1 Introduzione.....	5
1.1 I tipi di Sordità.....	5
1.1.1 Sindrome di Pendred e ipoacusia di tipo DFNB4.....	5
1.1.2 La componente genetica.....	6
1.2 La Pendrina.....	8
1.2.1 Super famiglia Sulfate Permeasi (SulP).....	8
1.2.2 La famiglia solute-linked carrier (SLC26).....	9
1.2.3 Funzioni di Pendrina.....	12
1.2.4 Dominio dei trasportatori del Solfato (SUL1).....	15
1.2.5 Dominio STAS (sulfate transporter - antisigma factor antagonist).....	18
1.2.6 Struttura quaternaria dei membri SLC26.....	20
1.3 Scopo della tesi.....	21
2 Materiali e Metodi.....	23
2.1 Analisi di sequenza e filogenesi.....	23
2.1.1 Ricerca di sequenze in banche dati.....	23
2.1.2 Allineamento di sequenze.....	24
2.1.3 Programmi per la visualizzazione di allineamenti multipli.....	24
2.1.4 Filogenesi.....	25
2.2 Analisi strutturale di regioni globulari.....	25
2.2.1 Metodi predittivi 1D.....	26
2.2.2 Banche dati di strutture proteiche.....	28
2.2.3 Programmi di visualizzazione di struttura terziaria.....	30
2.2.4 Predizione 3D.....	30
2.3 Analisi strutturale di regioni transmembrana.....	31
2.3.1 Composizione amminoacidica.....	32
2.3.2 Predizione della topologia.....	32
2.3.3 Predizione 2,5D.....	34
2.3.4 Predizione di Regioni rientranti.....	35
2.3.5 Predizione dell'orientamento delle eliche nella membrana.....	36
2.3.6 Presenza di proline-induced-kinks.....	38
2.3.7 Presenza di motivi GxxxG.....	38
2.4 Analisi dell'architettura funzionale delle proteine.....	39
2.4.1 Architettura dei domini della proteina.....	40
2.4.2 Identificazione di motivi lineari.....	40
2.4.3 Studio di residui discriminanti per la funzione.....	41
3 Risultati.....	43
3.1 Pendrina e la filogenesi della famiglia dei trasportatori SLC26A.....	44
3.2 Estremità N-terminale.....	48
3.3 Regione transmembrana.....	49
3.3.1 Predizioni 1D.....	49
3.3.2 Predizione 3D.....	58
3.4 Estremità C-terminale.....	59
3.4.1 Predizioni 1D.....	60
3.4.2 Predizione 3D.....	63
3.5 Analisi delle mutazioni.....	68

4	Discussione	79
5	Conclusioni	89
	Bibliografia.....	91

Abstract

Mutazioni del gene SLC26A4, codificante Pendrina, causano la Sindrome di Pendred e ipoacusia di tipo DFNB4. Più di 150 mutazioni sono state riportate ma non è ancora stato possibile individuare una correlazione genotipo-fenotipo. E' stato utilizzato un approccio integrato che combina informazioni biologiche, dati clinici e risultati ottenuti da metodi bioinformatici per individuare relazioni struttura-funzione e basi molecolari del meccanismo patogenetico.

E' stata rilevata la presenza di tre motivi conservati: il pattern conservato e il dominio transmembrana dei trasportatori SLC26, e il dominio STAS. L'analisi della topologia di Pendrina risulta nella predizione di 12 regioni transmembrana, tra cui una potenziale regione rientrante coinvolta nella selettività anionica dei trasportatori. L'identificazione di motivi GXXXG nell'ultima elica transmembrana e di struttura coiled-coil al C-terminale, potenziali siti di dimerizzazione, forniscono un supporto per il modello proposto di struttura quaternaria dimerica dei trasportatori SLC26. La struttura 3D del dominio STAS è stata predetta mediante homology modelling e l'analisi filogenetica sulla sequenza del dominio STAS ha permesso di individuare una caratteristica strutturale che differisce tra i paraloghi di Pendrina. La caratterizzazione in silico di Pendrina ha portato a un modello strutturale in cui la maggior parte delle mutazioni sono state mappate. Insieme alle nuove ipotesi che elucidano le relazioni struttura-funzione, è stata proposta una spiegazione degli effetti molecolari delle sostituzioni aminoacidiche e il loro possibile ruolo patogenetico.

1. Introduzione

1.1 I tipi di Sordità

Circa il 50% delle forme di ipoacusia infantile ha un'eziologia genetica, la maggioranza delle quali non è associata con anomalie di altri organi. Stime derivanti da dati clinici ed epidemiologici suggeriscono che l'80-85% delle forme di ipoacusia prelinguale non sindromica è a trasmissione autosomica recessiva, il 15% è autosomica dominante, e una piccola percentuale è ereditata come tratto X-linked o per trasmissione materna. Si conoscono circa 30 distinti loci (noti come loci DFNB) in cui la presenza di mutazioni può causare sordità non sindromica di tipo recessivo (NRSD). Nel locus DFNB4 è stato mappato il gene SLC26A4 responsabile di una forma di sordità sindromica caratterizzata da ipoacusia neurosensoriale associata a gozzo (Sindrome di Pendred, PS). La sindrome di Pendred sostiene il 10% delle ipoacusie ereditarie dell'uomo ed è per questo considerata la più comune forma di sordità sindromica.

1.1.1 Sindrome di Pendred e ipoacusia di tipo DFNB4

Mutazioni del gene SLC26A4, che codifica per il trasportatore anionico chiamato Pendrina, sono state descritte in soggetti con Sindrome di Pendred e soggetti con sordità non sindromica di tipo DFNB4 o EVA non-sindromica. Entrambe le patologie sono caratterizzate da ipoacusia neurosensoriale, di grado da severo a profondo, che insorge generalmente in età prelinguale, e possono essere associate ad anomalie dell'orecchio interno che variano dall'allargamento dell'acquedotto vestibolare (EVA) (Fig. 1) a una più complessa malformazione, che include l'ipoplasia cocleare, nota come Displasia di Mondini. In questa situazione, la sordità è comunemente congenita ma può anche precipitare a causa di traumi alla testa nella seconda decade di vita determinando un andamento fluttuante nella perdita di udito. Nella Sindrome di Pendred, il difetto nel trasporto di iodio può comportare la formazione del Gozzo, ma pressoché tutti i soggetti con tiroidomegalia sono clinicamente eutiroidi e hanno livelli sierici normali di T3 e T4. Sindrome di Pendred e ipoacusia neurosensoriale DFNB4 sono distinte clinicamente dall'associazione di gozzo eutiroidico o ipotiroidico in quanto le malformazioni dell'orecchio interno sono comuni ad entrambe le patologie. Purtroppo diversi studi riportano

una variabilità fenotipica intrafamiliar e interfamiliar, non solo nel grado di ipoacusia, ma soprattutto in termini di presenza/assenza e serietà delle malformazioni della tiroide e dell'orecchio interno. Inoltre il difetto tiroideo insorge classicamente nella seconda decade di vita per cui non è sempre possibile stabilire una diagnosi differenziale su basi esclusivamente cliniche.

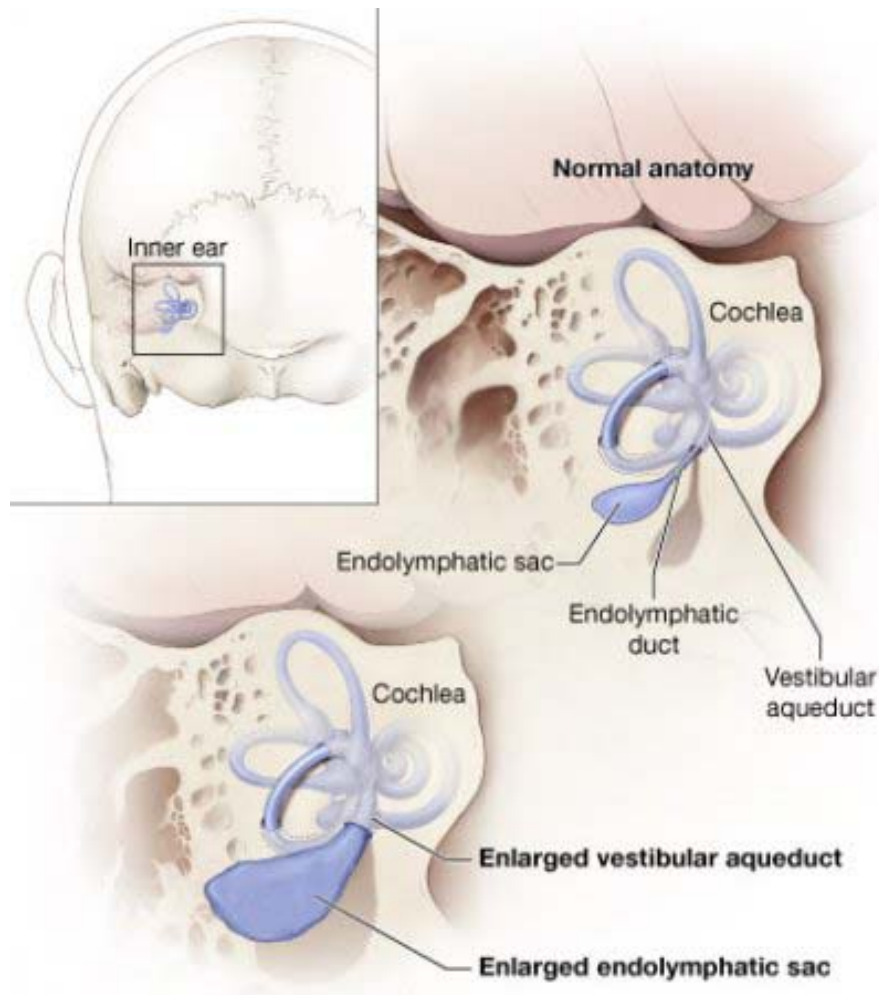


Fig. 1 Diagramma dell'orecchio interno. L'acquedotto vestibolare è un canale osseo che si allunga dall'orecchio interno alla scatola cranica, l'acquedotto inizia dall'osso temporale e prosegue con il dotto endolinfatico per finire nel sacco endolinfatico. L'acquedotto vestibolare si considera allargato quando al test con Risonanza Magnetica (MRI) o tomografia computerizzata (CT scan) il suo diametro appare maggiore di 1,5mm di larghezza.

(www.nidcd.nih.gov/health/hearing/eva.htm)

1.1.2 La componente genetica

Ad oggi circa 150 mutazioni SLC26A4 sono state identificate, incluse mutazioni missense e non sense, piccole delezioni e inserzioni e mutazioni nei siti di splicing. Recentemente sono state identificate

anche due grosse delezioni che comportano la perdita di una parte del gene SLC26A4. Le mutazioni non sembrano localizzare in particolari regioni hot-spot della proteina, ma sembra che la maggior parte coinvolga le estremità N e C terminale.

La frequenza delle mutazioni in diverse popolazioni è costante anche se si presentano con uno spettro diverso in base al gruppo etnico considerato. Dal punto di vista epidemiologico quindi le mutazioni SLC26A4 sembrano derivare da multipli e nuovi eventi mutazionali piuttosto che dalla diffusione di un antico allele fondatore (Park H-J., 2003).

E' stato osservato che alcune mutazioni sono state riscontrate soltanto in famiglie con sordità DFNB4, suggerendo una possibile correlazione genotipo-fenotipo. Si è ipotizzato che una funzione residua di Pendrina porti a un difetto biochimico che a livello della tiroide possa essere compensato da altri meccanismi, ma che a livello dell'orecchio interno risulti insufficiente per prevenire lo sviluppo di un difetto neurosensoriale. A sostegno di questa ipotesi, l'analisi funzionale, eseguita su alcune delle mutazioni che sono state associate alla Sindrome di Pendred, ha permesso di dimostrare la completa perdita dell'attività di trasporto di Pendrina, che viene invece preservata, anche se a livelli inferiori rispetto alla proteina wild-type, in presenza di mutazioni identificate solo in famiglie con EVA non-sindromica.

Altre mutazioni sono associate ad entrambi i tipi di patologia e soggetti con stesso genotipo presentano variabilità fenotipica intrafamiliare e interfamiliare. Recentemente è stato osservato che la presenza di due alleli mutanti SLC26A4 in soggetti con Sindrome di Pendred è significativamente più frequente rispetto a quella rilevata in soggetti con sordità DFNB4, in cui la percentuale di singoli alleli mutati risulta elevata. Alcuni studi, indicano l'EVA non-sindromica una condizione geneticamente eterogenea che è spesso causata da singole mutazioni in SLC26A4 in combinazione con altri fattori genetici o ambientali. Recentemente infatti è stato proposto un nuovo modello di meccanismo dosaggio-dipendente per la patogenesi molecolare di PS ed EVA non-sindromica, che sostiene il coinvolgimento non solo di SLC26A4 ma anche del suo apparato di controllo trascrizionale. E' stato dimostrato infatti che EVA non-sindromica segrega con eredità digenica, in un paziente portatore di una mutazione nel gene SLC26A4 e un'altra nel gene FOXI1, che codifica per un attivatore trascrizionale di SLC26A4 (T. Yang, 2007).

La varietà di dati disponibili fa sì che sia ancora aperto il dibattito su come considerare i due disordini e sul ruolo di Pendrina nel determinare l'insorgere di una o l'altra patologia. Alcuni sostengono l'idea di un *continuum* di patologia causata da mutazioni nello stesso gene (Suzuki H., 2007) mentre altri

appoggiano il concetto storico di PS e EVA non-sindromica, considerate come due distinte categorie di patologia.

1.2 La Pendrina

Pendrina presenta due domini principali identificati come dominio dei trasportatori del solfato (SUL1) e il dominio STAS (Sulfate transporter and antisigma-factor antagonist). La presenza del dominio dei trasportatori del Solfato classifica Pendrina nella famiglia delle Solfato Permeasi (SulP) e relativi trasportatori (SUL1) incluse nella Superfamiglia di Facilitatori. Il dominio STAS, localizzato nell'estremità citosolica carbossiterminale, è stato identificato mediante analisi di sequenza in proteine con funzioni completamente differenti come i trasportatori SulP e gli antagonisti batterici del fattore antisigma (ASA). ASA è un componente chiave nel meccanismo di sporulazione in risposta a condizioni di limitato apporto di nutrienti.

1.2.1 Super famiglia Sulfate Permeasi (SulP)

Pendrina è un membro della famiglia solute-linked carrier (SLC26) che fanno parte della Superfamiglia SulP (Sulfate Permease). Questa include permeasi per diversi substrati come xantina, uracile e vitamina C e comprende circa 200 membri derivanti da archae, batteri, funghi, piante e animali. Organismi quali: *Bacillus subtilis*, *Synechocystis sp*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* e *Caenorhabditis elegans* possiedono diversi paraloghi della famiglia SulP, la maggior parte codificanti trasportatori o scambiatori di anioni inorganici che differiscono per affinità verso il substrato e per meccanismo di trasporto. Alcune proteine funzionano per simporto $\text{SO}_4^{2-}:\text{H}^+$, ma per diversi omologhi è stato riportato un meccanismo ad antiporto $\text{SO}_4^{2-}:\text{HCO}_3^-$ o più generalmente anione:anione, mentre altri sono canali anionici. Fa eccezione la Prestina che agisce da proteina motore legata alla membrana nelle cellule ciliate esterne dell'orecchio interno responsabile del processo di amplificazione dello stimolo sonoro.

L'albero filogenetico della famiglia SulP rivela 5 rami o cluster principali. Tre di questi sono specifici dei batteri, gli altri due portano proteine eucariotiche con le proteine animali localizzate su un ramo e quelle di piante, lieviti e funghi sull'altro. Le proteine vegetali clusterizzano insieme, separate da quelle di lieviti e funghi. La presenza di paraloghi batterici distanti suggerisce che queste famiglie derivino in parte da eventi genici di duplicazione che

avvengono prima della divergenza tra i principali regni. Le proteine eucariotiche sono per la maggior parte più lunghe di quelle procariotiche. Le proteine batteriche variano in lunghezza da 434 a 566 residui, mentre le proteine eucariotiche variano in lunghezza da 611 a 893 residui.

Uno degli omologhi distanti di SulP è stato dimostrato essere uno scambiatore a simporto di $\text{HCO}_3^-:\text{Na}^+$ mentre i trasportatori del solfato sono Na^+ indipendenti. Studi di bioinformatica hanno permesso di identificare proteine che contengono il dominio SulP fuso con altri domini. Alcune di queste proteine presentano una combinazione di dominio SulP con dominio dell'anidrasi carbonica e dominio SulP con dominio della Rhodanasi, che sembrano essere rispettivamente coinvolti nell'uptake del bicarbonato e del solfato.

1.2.2 La famiglia solute-linked carrier (SLC26)

Ad oggi sono stati clonati 10 geni della famiglia solute-linked carrier (SLC26) che codificano per proteine capaci di trasportare una varietà di anioni monovalenti e divalenti, strutturalmente relati ai trasportatori del solfato ad alta affinità di piante e lieviti.

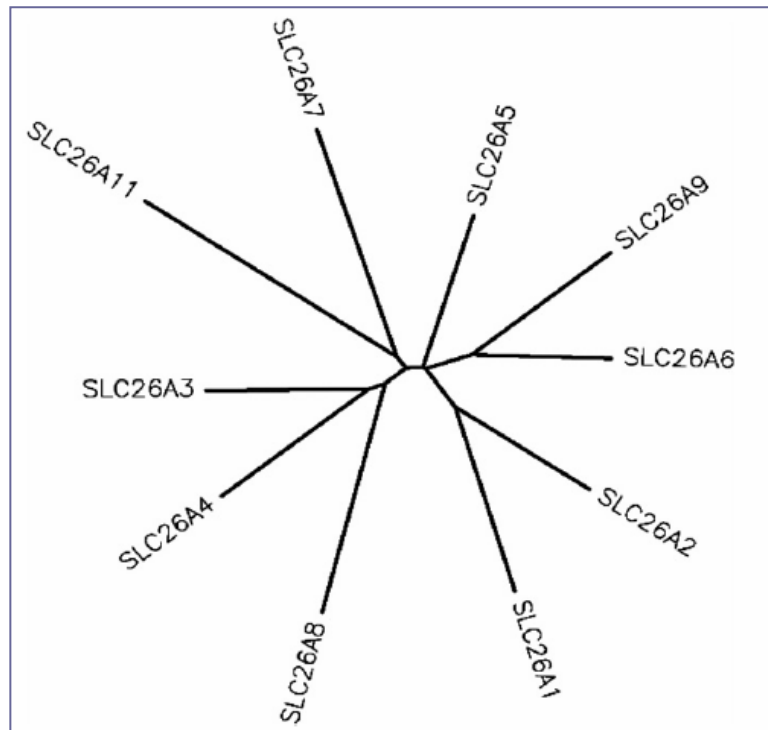


Fig. 2 Dendrogramma fenotipico che illustra la relazione tra i membri della famiglia SLC26 (Soleimani et al., 2006).

Una caratteristica delle proteine SLC26 è la relativamente bassa conservazione tra ortologi nel topo e nell'uomo tra cui la percentuale di identità varia tra il 76% a il 95%. Inoltre omologi di SLC26 in *Drosophila* e *C. elegans* mostrano un'identità del 25-40% con le proteine dei mammiferi. Questo implica che sia difficile riconoscere ortologi di ciascun membro SLC26 dei mammiferi. Dall'albero fenotipico senza radice, che illustra la relazione tra gli scambiatori anionici SLC26 umani, si nota che c'è una stretta omologia tra alcuni di questi membri. SLC26A1 e SLC26A2 mostrano il maggior grado di omologia tra di loro mentre SLC26A3 è più vicino a SLC26A4. SLC26A6 presenta maggiore similarità di sequenza con SLC26A9, e SLC26A7 con SLC26A11 (Fig. 2).

Per l'identificazione di ortologi di SLC26A5, recentemente è stata applicata una più complessa strategia che utilizza informazioni aggiuntive alla similarità di sequenza. L'approccio ha stabilito i seguenti criteri: 1. la conservazione del *linkage* cromosomico; 2. similarità nell'organizzazione genica nella disposizione degli esoni e nei confini esone-introne; 3. similarità a livello nucleotidico e aminoacidico; 4. conservazione di motivi strutturali nei domini SUL e STAS. Grazie a questo studio è stata stabilita l'evoluzione della famiglia SLC26A e la conservazione della sequenza proteica di pertinenza all'interno dei mammiferi ma sono anche emerse particolari caratteristiche strutturali dei trasportatori del solfato.

I geni della famiglia SLC26A sono stati divisi in due principali gruppi, nominati SLC26A4/11 (9 membri) e SLC26A5/6 (2 membri). Il gruppo SLC26A4/11 può essere suddiviso in due subfamiglie in base alla similarità di sequenza e all'organizzazione genomica. La subfamiglia SLC26A4 è rappresentata dal gene *slc26a4* come riferimento e include i membri *a3*, *a4*, *a8*, *a9*, mentre la subfamiglia SLC26A11 contiene il prototipo *Slc26a11* e include i geni *a1*, *a2* e *a10*. Come visto in altre famiglie di geni, il numero e le differenze nelle sequenze di paraloghi appare espandersi e contrarsi con l'evoluzione delle diverse specie eucariote. Nelle alghe, nei vermi e negli insetti i membri della famiglia SLC26A mostrano un numero di esoni inferiore rispetto ai vertebrati. Dal punto di vista della proteina sono state identificate tre regioni, che mostrano una certa variabilità di sequenza tra gli euteria e i vertebrati non-placentati, localizzate alle estremità N e C terminale e in una regione ricca di residui carichi del dominio carbossiterminale. Le variazioni di sequenza osservate tra i due gruppi di organismi sono dovute a delezioni o inserzioni a livello di queste regioni. In particolare, delezioni all'interno della sequenza codificante dell'esone 4 comportano la riduzione in lunghezza di un loop idrofilico e inserzioni all'interno dell'esone 6 aggiungono un' α - elica idrofobica probabilmente transmembrana. Accanto alle sostituzioni di singoli

residui tra ciascun paralogo, sono state osservate inserzioni/delezioni all'interno degli esoni, mentre i confini esone/introne rimangono conservati. Lo studio della sequenza di Prestina nei vertebrati ha individuato la presenza di Indel in 5 esoni codificanti: 1, 4, 6, 16 e 18. L'1 e il 18 rappresentano le estremità della proteina, gli esoni 4 e 6 codificano per regioni interne al dominio SLP mentre l'esone 16 rappresenta il cluster di residui carichi nel dominio carbossiterminale. Le delezioni/inserzioni che interessano gli esoni 4 e 6 sembrano essere un'innovazione dei mammiferi. Esse infatti determinano il principale cambiamento che permette il passo finale nella formazione di un motivo essenziale minimo della sequenza di Prestina nei mammiferi (O. E. Okoruwa et al., 2008).

Gene	Locus	Esoni	Proteina	(aa)	Substrato	Distribuzione tissutale	Malattia
SLC26A1	4p16.3	3	Sat-1	701	SO ₄ ²⁻ , ossalato, HCO ₃ ⁻	Fegato, rene. Bassi livelli in cervello e polmone	
SLC26A2	5q31-34	4	DTDST	739	SO ₄ ²⁻ , Cl ⁻	Osteoblasti, timo, testicoli, rene	Condrodisplasia
SLC26A3	7q31	21	DRA, CLD		SO ₄ ²⁻ , Cl ⁻ , HCO ₃ ⁻ , OH ⁻ , ossalato	Intestino, ghiandole sudoripare, pancreas, prostata	Cloridrorrea congenita
SLC26A4	7q31	21	Pendrin	780	Cl ⁻ , HCO ₃ ⁻ , I ⁻ , formiato, fruttosio e mannosio	Orecchio interno, rene, tiroide	Sindrome di Pendred e sordità (DFNB4)
SLC26A5	7q22	20	Prestina		?	Orecchio interno, cervelli, testicoli e tessuto vestibolare	Sordità non sindromica
SLC26A6	3p21	21	CFEX, PAT-1	738	Cl ⁻ , HCO ₃ ⁻ , OH ⁻ , ossalato, formiato	Rene, intestino, cuore, cervello, muscolo scheletrico, polmone	Calcio ossalato nefrolitiasi
SLC26A7 2 varianti A e B	8q22.2	19	PAT-2	A: 656	SO ₄ ²⁻ bassa affinità, Cl ⁻ , ossalato, HCO ₃ ⁻	Stomaco, Rene e placenta. Bassi livelli nel testicolo	
SLC26A8	6p21.3	22	TAT-1		SO ₄ ²⁻ , Cl ⁻ , ossalato	Testicoli, cervello, rene	
SLC26A9	1q32	21	PAT-4		SO ₄ ²⁻ , Cl ⁻ , ossalato	polmoni	
SLC26A10	12q13		/		Pseudogene	/	
SLC26A11 Esistono varianti	17q25	18 o 17	PAT-5		SO ₄ ²⁻ , Cl ⁻ , HCO ₃ ⁻	Placenta, cervello, rene, intestino	

Tabella 1: Riassunto della famiglia di geni SLC26A.

Proprietà dei membri SLC26A

Alcuni membri della famiglia SLC26 funzionano come trasportatori del Solfato ma altri, come SLC26A4/Pendrina e SLC26A6/PAT-1, possono trasportare diversi anioni ma non il solfato. Inoltre uno dei

membri SLC26, SLC26A5/Prestina appare funzionare come proteina motore delle cellule ciliate esterne che cambia la sua conformazione utilizzando gli anioni intracellulari come estrinseco sensore del voltaggio.

Il ruolo fisiologico dei diversi paraloghi è evidentemente dovuto alla variazione nella specificità dell'anione trasportato e nel pattern di espressione (D. B. Mount et al., 2004). I membri di questa famiglia sono espressi nella membrana apicale e basale di diversi tessuti e sono principalmente coinvolti nel trasporto vettoriale di vari anioni. Uno specifico interesse nella famiglia SLC26 è stato ulteriormente stimolato dal fatto che 4 geni SLC26 umani sono associati con patologie recessive fenotipicamente distinte (vedi Tab.1).

1.2.3 Funzioni di Pendrina

L'inusuale combinazione di sordità e difetto tiroideo incontrato nella Sindrome di Pendred fa nascere alcune interessanti questioni circa il modo in cui difetti della stessa proteina portano a distinti effetti tessuto-specifici. Studi di espressione e di sublocalizzazione cellulare hanno permesso di ipotizzare il ruolo di questa proteina nei diversi tessuti in cui è presente (Royaux et al., 2000)

Pendrina è espressa nell'orecchio interno, nella tiroide, nel rene, nella mammella e nei testicoli. Espressione di Pendrina è stata rilevata anche nell'endometrio, dove sembra avere una localizzazione differente durante il ciclo mestruale, e nella placenta, dove l'espressione cresce durante la gestazione. Nel topo i livelli di espressione di Pendrina sono maggiori nel rene rispetto alla tiroide. Questo è in linea con l'assenza di alterazioni tiroidee nel topo KO per il gene SLC26A4. Si suppone che ci sia una differenza tra topo e uomo nella funzione stessa di Pendrina, ma è più probabile che siano coinvolti diversi fattori di regolazione che possono influenzare l'espressione di Pendrina. La sequenza di SLC26A4 infatti differisce sostanzialmente da quella di topo a livello del 3' UTR.

Per quanto riguarda l'attività di trasporto, Pendrina sembra avere funzioni distinte nei diversi tessuti in cui è espressa. Nonostante diversi studi funzionali abbiano permesso di dimostrare che Pendrina è importante per il trasporto di Cl^- , HCO_3^- , I^- , formiato, ma anche di zuccheri come fruttosio e mannosio, il meccanismo elettrofisiologico del trasporto di anioni non è ancora chiaro. La maggior parte degli studi concorda nell'affermare che Pendrina funziona come uno scambiatore anionico che generalmente trasporta il Cloro all'interno della cellula attraverso lo scambio con altri anioni in modo dipendente dal potenziale di membrana. La misura della corrente uscente e entrante indica che Pendrina è importante per il riassorbimento di Cloro e l'efflusso di Iodio, anche se lo scambio di

Iodio nel medium con il Cloro citoplasmatico avviene a concentrazioni intracellulari di Iodio significativamente elevate. Iodio e Cloro inoltre sono richiesti in entrambi i lati della membrana perché avvenga l'attività di scambio, indicando che gli anioni stessi attivano la funzione di Pendrina mentre sono trasportati.

Ruolo fisiologico di Pendrina nell'orecchio interno

L'espressione di pendrina è stata dimostrata nel dotto (ED) e nel sacco endolinfatico (ES) e in altre regioni ristrette dell'orecchio interno, come le cellule della prominenza spirale e della *stria vascularis*. Il dotto endolinfatico è la parte del labirinto membranoso che connette la coclea e l'apparato vestibolare al sacco endolinfatico, localizzato nella fossa craniale posteriore. Questi due organi sono suddivisi in due differenti compartimenti riempiti da fluidi a diversa composizione: la perilinfia simile ai fluidi extracellulari e l'endolinfa ricca in potassio e proteine, e povera di sodio. L'endolinfa presenta un potenziale positivo (+ 80mV) comparato con la perilinfia o il plasma.

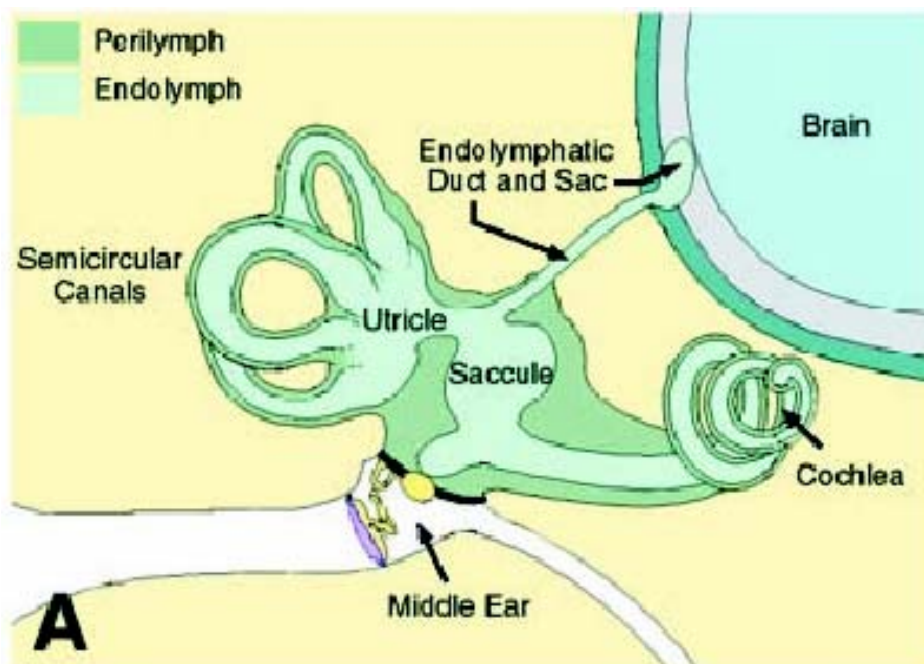


Fig. 3 Compartimenti nella coclea e regioni coinvolte nel mantenimento dei fluidi endococleari.

Il caratteristico pattern di espressione di Pendrina ha fornito le basi per ipotizzare il suo ruolo nel mantenimento dell'omeostasi dell'endolinfa (Everett et al., 1999). Nel dotto cocleare, l'endolinfa è secreta principalmente dalle cellule marginali della *stria vascularis*. Inoltre il sacco endolinfatico, area di maggiore espressione di

1. Introduzione

Pendrina, presenta un particolare tipo di cellule, con caratteristiche ultrastrutturali simili alle cellule intercalate del rene, che sembra avere un ruolo cruciale nel processo di riassorbimento dell'endolinfa. In particolare, il subtipo B delle cellule ricche in mitocondri, funzionano come scambiatori di Cl-/bicarbonato e vengono attivate da una diminuzione del volume endolinfatico. Dovrebbero rappresentare quindi la sede in cui ha inizio il meccanismo patogenetico che sostiene il difetto di udito. E' stato ipotizzato che la perdita di funzione di pendrina a questo livello può risultare in un progressivo aumento di volume a cui segue un allargamento del labirinto membranoso e delle strutture ossee che lo contengono. Il danno provocato all'epitelio porta a un difetto di udito neurosensoriale. Questa condizione è simile a quella osservata in topi KO per SLC26A4, in cui il volume dell'endolinfa è aumentato e c'è una perdita del potenziale endococleare generato da canali del potassio localizzati nelle cellule intermedie. E' stato dimostrato che un difetto di Pendrina comporta una perdita dei canali del Potassio KCNJ10 indotta dall'aumento di stress ossidativo conseguente all'aumento di fluidi endolinfatici e all'ipercidificazione dell'endolinfa.

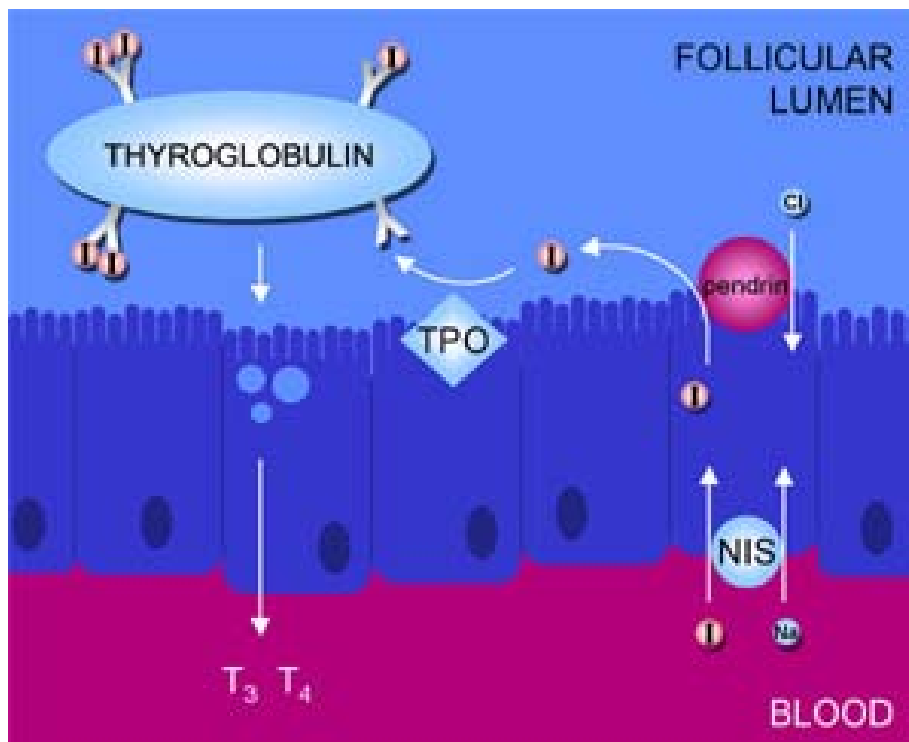


Fig. 4 Passaggi chiave della organificazione dello iodio. Il passaggio dello Iodio dal tirocita alla colloide è regolato da Pendrina.

Nella Tiroide

Nella tiroide, Pendrina è espressa nella membrana apicale dei tirociti, dove regola il flusso di iodio nel follicolo tiroideo. Lo Iodio viene assorbito dai capillari che rivestono i follicoli tiroidei grazie al trasportatore a simporto di iodio e sodio (NIS) localizzato nella membrana basale dei tirociti, da cui viene trasportato alla membrana apicale per poi essere organificato nella colloide. In questo compartimento lo Iodio viene legato a residui di Tirosina per formare la tiroglobulina, che darà origine agli ormoni T3 e T4. Il passaggio dello Iodio dal tirocita alla colloide è regolato da Pendrina. Nei soggetti con sindrome di Pendred si osserva, in seguito alla somministrazione di Perclorato, un inappropriato rilascio di Iodio dai tirociti, in forma di tiroglobulina. Il test al Perclorato (PDT) mostra però solo un parziale difetto di organificazione (PIOD) in accordo con i fenotipi non gravi, spesso eutiroidei, riscontrati nei pazienti. Questo ha fatto supporre che, in assenza della funzione di pendrina, il flusso di Iodio nella colloide possa avvenire attraverso uno o più sistemi di trasporto.

Nel rene

L'RNA messaggero di Pendrina è stato rilevato a livello dei dotti collettori corticali del rene e l'analisi immunoistochimica ha permesso di individuare che Pendrina è specificatamente localizzata nella membrana apicale delle cellule intercalate B e nonA-nonB. Nel rene, Pendrina sembra avere un ruolo nell'omeostasi acido-base sistemica mediante la secrezione di Bicarbonato nei tubuli collettori. Si è visto infatti che la sua espressione diminuisce in condizioni di ipocaliemia e aumenta in condizioni di alcalosi metabolica. Recentemente, inoltre, è stato studiato un possibile ruolo di Pendrina nell'ipertensione mediata da mineralcorticoidi in risposta a un limitato apporto di NaCl. In queste condizioni SLC26A4 viene sovraregolata per poter mantenere l'equilibrio acido-base e la conservazione di Cl⁻ e acqua. La sua assenza comporta un difetto di assorbimento di NaCl dai tubuli collettori che risulta in ipertensione.

1.2.4 Dominio dei trasportatori del Solfato (SUL1)

Il dominio SUL1, altamente conservato nella famiglia di geni SLC26A, corrisponde al dominio di membrana deputato al passaggio di anioni. Esistono poche informazioni riguardo la struttura secondaria e terziaria di questo dominio e rimane ancora dibattuto l'esatto numero di segmenti transmembrana e la loro topologia. Per alcuni

trasportatori SLC26, numerose strategie sono state utilizzate allo scopo di studiare la relazione struttura-funzione e in alcuni casi hanno permesso di individuare specifici residui o domini coinvolti nella funzione degli scambiatori anionici. L'analisi mutazionale è stata indirizzata a residui carichi, siti di fosforilazione, siti di glicosilazione e residui di cisteina. Per alcune proteine sono stati analizzati gli effetti funzionali e strutturali determinati dalla delezione delle estremità della proteina o dalla sostituzione aminoacidica a livello di siti di mutazione associati alla relativa malattia.

Motivo conservato dei trasportatori del solfato

Le prime eliche transmembrana contengono una sequenza di omologia, caratteristica della famiglia SLC26, di 22 aminoacidi inizialmente definita per confronto del primo membro SLC26 di mammiferi con omologhi negli organismi inferiori. Si presume che queste tre eliche siano strettamente accoppiate in quanto connesse da *loop* molto corti. Studi di mutagenesi sitospecifica in SHST1, un trasportatore del solfato di *Stylosanthes hamat*, hanno permesso di dimostrare l'importanza di tre residui di Prolina nella conformazione delle tre eliche accoppiate che sembrano essere cruciali per l'attività del trasportatore (Shelden, 2001). La prolina gioca un ruolo funzionale nelle proteine di membrana, dove è presente con una frequenza inaspettata. L'incorporazione di una prolina in un α -elica porta alla perdita di un legame idrogeno nella catena carboniosa, a causa dell'impedimento sterico con il residuo successivo, causando una distorsione della normale geometria dell'elica. Queste distorsioni possono creare dei punti deboli nell'elica che sembrano facilitare il movimento richiesto per la funzione di alcune proteine canale (Yohannan et al., 2004).

Analisi *Evolutionary Trace* (ET) raggruppa residui correlando pattern di variazioni aminoacidiche con la divergenza funzionale nelle famiglie proteiche. L'applicazione del metodo alla Prestina ha permesso di individuare due *cluster* di residui altamente ravvicinati nella struttura secondaria e terziaria all'interno del motivo dei trasportatori del solfato. Mutagenesi sito specifica su alcuni di questi residui mostra che soltanto sostituzioni con aminoacidi di diversa grandezza influenzano la capacità di scambiare anioni della proteina suggerendo che tali residui siano coinvolti nelle interazioni tra le eliche e influenzino il cambiamento conformazionale associato al movimento delle cariche. Questa regione potrebbe rappresentare la tasca di legame del Cloro che si pensa funzioni come sensore del voltaggio. Diversi trasportatori e anioni infatti indicano che il cloro è coordinato mediante una parziale carica positiva dalla catena principale di un amide e la selettività è

conferita dalle dimensioni del filtro e non dalla carica (Rajagopalan et al., 2006).

Motivo di Saier

Esiste un secondo cluster di residui invariabili al C-terminale del core idrofobico della proteina, ottenuto dal multiallineamento di sequenze proteiche della Superfamiglia SulP costruito da Saier et al. (motivo di Saier). Questa regione include la tripletta NQE in cui un unico residuo (E384 in Pendrina) risulta completamente conservato in tutte le proteine della Famiglia SulP (Sayer et al., 1999). La rilevanza di questo sito nel folding e nella funzione di Pendrina è stata indagata in diversi studi di mutagenesi e risulta che E384 sia importante per l'integrità strutturale di Pendrina e che un'alterazione a livello di questo sito provochi un difetto di folding. La proteina mutante viene infatti trattenuta nel reticolo endoplasmatico con l'impossibilità di raggiungere la membrana (Rotman-Pikielny, 2002; Yoon et al., 2008).

Siti di glicosilazione

Il processo di N-glicosilazione è una modificazione della proteina che avviene durante la traduzione e in molti casi è determinante per la localizzazione di proteine di membrana. Potenziali siti di N-glicosilazione sono presenti in tutte le sequenze proteiche SLC26 di mammifero tra la terza e la quarta elica transmembrana e la loro localizzazione è conservata. Nelle sequenze dei trasportatori di funghi e piante sono invece assenti e, in corrispondenza di questa regione, l'allineamento introduce un gap (Smith et al., 1995).

Per alcuni membri della famiglia SLC26 lo stato di glicosilazione è stato dimostrato ma per altri, come per Pendrina, non sono ancora stati identificati i siti di glicosilazione. La proteina SLC26A3/DRA è espressa nell'intestino in tre forme eterogenee che differiscono per lo stato di N-glicosilazione (Byeon et al., 1998).

Prestina è una glicoproteina con N-glicosilazione ai residui N163 e N166, quest'ultimo sembra essere il sito preferenzialmente glicosilato. La proteina completamente deglicosilata mostra un'alterata funzione elettrofisiologia ma lo stato di glicosilazione sembra non essere necessario per il targeting alla membrana di Prestina. Esperimenti di immunoblotting rilevano la presenza di due bande specifiche per Pendrina di 85 e 110 KDa, un peso molecolare superiore a quello calcolato di 82 KDa. Il trattamento con peptide N-glicosidasi permette di ottenere una banda di 82 KDa indicando che la banda di 85KDa sia relativa a una forma del Reticolo endoplasmatico (RE), glicosilata all'interno del core della

proteina, mentre la banda a 110 KDa si riferisce a una forma glicosilata più complessa, che a differenza di Prestina, è richiesta per il *targeting* alla membrana plasmatica.

Residui di Cisteina

Il ruolo di residui di cisteina è stato indagato per alcuni trasportatori del solfato. In particolare sono state studiate delle varianti prive di cisterna di SHST1, un trasportatore del solfato di *Arabidopsis*, dimostrando che nessuna delle 5 cisteine presenti nella sequenza di SHST1 è essenziale per la funzione (Howitt, 2005).

1.2.5 Dominio STAS (sulfate transporter - antisigma factor antagonist)

Il dominio citoplasmatico C-terminale di tutte le proteine SLC26 include il dominio STAS (sulfate transporter - antisigma factor antagonist) scoperto grazie alla sua similarità di sequenza tra la famiglia SLC26 e antagonisti batterici del fattore anti-sigma (ASA), come SpollAA di *Bacillus subtilis*. SpollAA è un polipeptide che interagisce con il fattore anti-sigma SpollAB, lasciando libero il fattore sigma di regolare l'attività dell'RNA polimerasi nel processo di sporulazione. SpollAA lega GTP e ATP e possiede una debole attività NTPasica che viene abolita mediante fosforilazione da parte di SpollAB a livello di un residuo di serina localizzato in un motivo conservato. La conservazione di questo motivo anche nelle proteine SLC26 ha suggerito che il dominio STAS possa avere una attività di legame a fosfonucleotidi che regola l'attività di trasporto degli anioni (Aravind and Koonin, 2001). Fin'ora però in nessun membro SLC26 è stato dimostrato che la funzione di trasporto dipenda dalla concentrazione intracellulare di nucleotidi.

Funzione del dominio STAS

Il ruolo fisiologico o meccanicistico del dominio STAS negli scambiatori SLC26 non è noto ma sembra avere una funzione regolativa nell'attività di trasporto degli anioni. Il dominio STAS di alcuni membri SLC26 è stato studiato mediante mutagenesi sitospecifica, delezioni seriali del dominio C terminale e proteine chimeriche, costruite mediante la fusione della porzione C-terminale derivante da proteine omologhe della famiglia SLC26. In questo modo è stato possibile raccogliere informazioni riguardo il ruolo del dominio STAS nel controllo dell'attività degli scambiatori SLC26 e nella loro stabilità, o per la localizzazione alla membrana plasmatica. Mutanti deleti e proteine chimeriche mostrano perdita

di funzione della proteina e ritenzione nel RE e nell'apparato del Golgi probabilmente dovuta ad un inappropriato folding. Le delezioni seriali hanno dimostrato che alcune regioni sono indispensabili per il targeting della membrana mentre altre no e che non tutti i membri SLC26 necessitano del C-terminale per la loro localizzazione subcellulare e funzione. Sembra che l'espressione delle proteine a livello della membrana sia dipendente da sequenze di residui, specifici per ciascun membro SLC26, localizzati nel C-terminale. Motivi contenenti Tirosina (YXX Φ , dove Φ è idrofobico) e motivi di-leucina sono tra i motivi più studiati che regolano il trasporto tra l'apparato del Golgi e la membrana plasmatica (Zheng et al., 2005). Queste sequenze potrebbero essere coinvolte nell'interazione intramolecolare del dominio STAS con il dominio di membrana o con interazioni intermolecolari con altre proteine o metaboliti cellulari (Shibagaki et al., 2004). Recentemente infatti è stato dimostrato che il dominio STAS di SLC26A3 (downregulated in adenoma DRA) interagisce, mediante proteine adattatrici PDZ, con il dominio R di CFTR (cystic fibrosis transmembrane conductance regulator). L'interazione risulta in una attivazione di entrambe le proteine (Soleimani et al., 2006).

Analisi della Struttura terziaria del dominio STAS

Il dominio STAS delle proteine ASA batteriche è stato strutturalmente caratterizzato mediante spettroscopia NMR e cristallografia ai raggi-X. SpolIAA possiede un α/β fold privo di lunghi loop e di regioni disordinate. Al centro del dominio si trova uno scaffold centrale costituito da quattro β -strand con il primo a orientamento antiparallelo rispetto agli altri. Un lato del foglietto- β è coperto completamente da 2 α -eliche mentre l'altro lato è coperto parzialmente dall'estremità C-terminale, dove si trova una substruttura a forma tipica di una mano contenente la terza α -elica (helical handle). Questa regione, ancorata al core idrofobico da residui aromatici (Kovacs et al., 1998), ha una posizione poco definita in confronto al resto della struttura, probabilmente determinata da una maggiore flessibilità di questa regione. Il loop tra il terzo β -strand e la seconda α -elica contiene il motivo DSSL altamente conservato in cui si trova il sito di fosforilazione.

A differenza delle proteine ASA batteriche, si conoscono poche informazioni riguardo la struttura del dominio STAS presente negli scambiatori SLC26. La difficoltà di ottenere una struttura 3D di tali domini deriva dal fatto che questi peptidi hanno una forte tendenza ad aggregare. Esperimenti di dicroismo circolare (CD), spettroscopia a fluorescenza su porzioni del C-terminale di Prestina, contenente il dominio STAS, confermano la presenza di una struttura

tridimensionale definita. Inoltre la struttura secondaria determinata sperimentalmente è in accordo con quella derivata dalla struttura 3D del dominio STAS batterico.

L'allineamento tra sequenze proteiche ASA batteriche e scambiatori anionici rivela una forte conservazione del loop contenente il motivo di fosforilazione, che nei membri SLC26 presenta due residui invarianti DxxL. Le differenze maggiori invece si osservano a livello del loop tra la prima α -elica e il terzo β -strand. Negli scambiatori SLC26 questo loop è un sito di inserzioni significative e può raggiungere la lunghezza di 150 residui nel caso di SLC26A8. Gli omologhi SLC26 batterici non presentano inserzioni e nei paraloghi di *Drosophila* e *C. elegans* il loop risulta anche più corto. Nei trasportatori SLC26 è presente inoltre una estensione del C-terminale che risulta variabile anche tra i diversi membri SLC26.

1.2.6 Struttura quaternaria dei membri SLC26

Recentemente sono state ottenute evidenze sperimentali che dimostrano che le proteine SLC26 aggregano a livello della membrana plasmatica, formando dimeri o tetrametri stabili che sembrano indispensabili per la funzione fisiologica della proteina.

Per Prestina era stata proposta inizialmente una struttura a tetramero costituita da due unità dimeriche connesse da un legame disolfuro. Anche esperimenti di FRET-analisi supportano la struttura oligomerica e la ricostruzione in 3D del tetramero appare a forma di pallone (Mio, 2008). Recentemente altri esperimenti indicano che i trasportatori SLC26 sono assemblati in dimeri, composti da due subunità identiche funzionanti in modo dipendente l'una dall'altra. Inoltre per quanto riguarda Prestina i dimeri sembrano interagire tramite legami non covalenti piuttosto che con ponti disolfuro (Detro-Dassen, 2007).

Porzioni di C-terminale contenente il dominio STAS sono state studiate per la loro propensione ad aggregarsi formando oligomeri e multimeri di alto peso molecolare sensibili alla composizione del tampone in cui sono immersi. Lo stato oligomerico è influenzato da specie ioniche e da sostanze apolari per cui si pensa che l'associazione sia mediata da interazioni sia polari che apolari. In particolare il tipo e la concentrazione dell'anione in soluzione influenza la formazione degli aggregati. L'aumento della concentrazione di NaCl da 10 a 100mM sposta l'equilibrio da una situazione in cui i dimeri prevalgono ad un'altra in cui il tetramero è la forma più abbondante. Le proprietà di aggregazione di questi peptidi sono interessanti perché potrebbero avere un ruolo nella regolazione della funzione della proteina nella sua lunghezza intera (Pasqualetto et al., 2008).

1.3 Scopo della tesi

Data la difficoltà di ottenere una struttura tridimensionale di proteine transmembrana, non esistono molte informazioni che permettano di costruire un modello strutturale di Pendrina. Fin'ora, per descrivere la proteina, in letteratura è sempre utilizzata una rappresentazione topologica molto semplice che fornisce poche informazioni relative alla relazione struttura-funzione. Lo scopo della presente tesi è dunque triplice.

In primo luogo si tratta di utilizzare i più avanzati metodi di analisi *in silico* per caratterizzare al meglio gli aspetti strutturali e funzionali della proteina al fine di costruire un modello strutturale. Questi dati dovrebbero servire a sviluppare un'ipotesi funzionale della proteina che ne spieghi il funzionamento a livello molecolare. Si intende poi utilizzare le informazioni strutturali e l'ipotesi funzionale per mappare ed analizzare le mutazioni di Pendrina che causano sindrome di Pendred o EVA non-sindromica. Lo scopo ultimo è quello di trovare un modello di correlazione genotipo-fenotipo che sia in grado di predire gli effetti delle mutazioni a livello molecolare.

2. Materiali e Metodi

La conoscenza della struttura tridimensionale delle proteine è essenziale per comprendere i dettagli della loro funzione molecolare e fornisce informazioni per lo sviluppo strategico di esperimenti per il disegno di nuovi farmaci o per lo studio di mutazioni relate a patologie. In questo ambito una crescente attenzione è stata data alla bioinformatica che permette di analizzare l'enorme quantità di dati biologici disponibili con sofisticate tecniche di analisi e vede un continuo sviluppo di metodi di predizione che consentono di accelerare il processo di scoperta e formulazione di nuove ipotesi. In questo lavoro metodi computazionali sono stati utilizzati per caratterizzare strutturalmente e funzionalmente una proteina di interesse biomedico. Le varie applicazioni, descritte in seguito, riguardano l'identificazione e l'allineamento di sequenze omologhe, la caratterizzazione dell'architettura primaria della proteina, che comprende domini e motivi di legame, e la predizione di struttura secondaria e terziaria della proteina. La localizzazione strutturale delle varianti di sequenza associate alla patologia ha permesso infine di dare una spiegazione alla perdita di funzione che ne deriva.

2.1 Analisi di sequenza e filogenesi

2.1.1 Ricerca di sequenze in banche dati

Le banche dati primarie più comunemente utilizzate, **EMBL** e **GenBank**, contengono sequenze di acidi nucleici e un minimo di informazione sulla sequenza. Le sequenze proteiche vengono raccolte nelle banche dati Uniprot (già TREMBL) e SWISSPROT. **SWISSPROT** è la banca dati di proteine di riferimento per tutti gli studi correlati ad analisi in silico di proteine e patterns proteici e presenta un'accurata annotazione del nome della proteina e del codice relativo al gene che la codifica. SWISSPROT è aggiornata dal gruppo svizzero in collaborazione con l'EBI (European Bioinformatics Institute) dove viene sviluppata l'altra banca dati di proteine, TREMBL, risultato della traduzione automatica in aminoacidi di tutte le sequenze annotate nella banca dati EMBL come sequenze codificanti proteine. Di queste sequenze annotate in TREMBL, una parte viene incamerata in SWISSPROT. La sequenza proteica di Pendrina è stata recuperata con *Entrez*, il sistema di interrogazione della banca dati GenBank, sviluppato dal NCBI (National Center for Biotechnology information).

Per la ricerca di sequenze omologhe è stato utilizzato BLAST o PSIBLAST con soglia di e-value 0.005 per inclusione nel profilo ed e-value 10 per la ricerca di sequenze più distanti. Il server **PSI-BLAST** utilizza una procedura iterativa per cui tutte le sequenze che superano la soglia imposta partecipano alla creazione di un modello detto PSSM (Position Specific Substitution Matrix) utilizzata nei successivi cicli per ricercare sequenze evolutivamente distanti. Le sequenze utilizzate nella costruzione della PSSM sono state controllate manualmente ed è stato necessario includere alcune sequenze che non superavano la soglia alla prima iterazione. Per evitare fenomeni di deriva sono state utilizzate al massimo 4 iterazioni.

PDB-BLAST è stato utilizzato come protocollo semplice di fold recognition. Utilizza il fatto che PSI-BLAST permette di salvare il PSSM e ripartire con una nuova ricerca usando lo stesso PSSM. Per massimizzare l'informazione su sequenze simili, si inizia una ricerca contro NR (non redundant; tutte le sequenze in UniProt) per 4 iterazioni, costruendo un PSSM che viene poi utilizzato per cercare strutture simili nel database PDBAA di sequenze con struttura nota.

2.1.2 Allineamento di sequenze

Gli allineamenti multipli di sequenze (MSA) sono stati assemblati con metodi basati sulla procedura di allineamento progressivo, come **CLUSTALW**. Questa metodica può incorrere in problemi di minimo locale, una volta che due o più sequenze sono state allineate in un cluster, questo allineamento resta fissato e non può più essere modificato nelle fasi successive dell'allineamento progressivo. L'allineamento multiplo risulta inoltre problematico quando le sequenze che si desidera allineare sono di lunghezza molto diversa per la presenza di grosse inserzioni o delezioni. Il programma **MAFFT**, che permette un miglior allineamento locale, è stato utilizzato nella fase di analisi della regione variabile in lunghezza del dominio STAS.

2.1.3 Programmi per la visualizzazione di allineamenti multipli

Per valutare la qualità di un multiallineamento si possono usare vari sistemi di visualizzazione che offrono oltre alla rappresentazione a colori la possibilità di modificare l'allineamento sulla base di informazioni accessorie derivanti da indagini sperimentali o strutturali, forzando l'allineamento sulla base della conservazione di elementi di struttura secondaria o terziaria. **Jalview**, un editor scritto in Java, è stato utilizzato per la visualizzazione di multiallineamenti e nel confronto dei risultati ottenuti dalle diverse predizioni per lo studio della regione transmembrana di Pendrina.

Esript è uno strumento che permette di presentare un multiallineamento arricchito di informazioni, provenienti da altri metodi di predizione, come la struttura secondaria, o di indicare particolari caratteristiche relative a specifiche posizioni dell'allineamento.

2.1.4 Filogenesi

L'analisi filogenetica sulla sequenza del dominio STAS dei trasportatori del solfato è stata eseguita utilizzando il programma PHYML 2.4 basato sul metodo di maximum likelihood (ML). La ricostruzione dell'albero filogenetico è stata effettuata mediante matrice di sostituzione JTT, mentre l'eterogeneità dei siti è stata modellata con una distribuzione Γ di quarta categoria. La topologia dell'albero, visualizzata con i programmi Treeview 1.6.6 e NJplot, è stata valutata mediante ricampionamento bootstrap nonparametrico. Come outgroup è stata utilizzata la sequenza di *Paracoccus denitrificans* scelta in base ai dati di letteratura. L'albero filogenetico derivante è sufficientemente complesso per rappresentare un modello di evoluzione che tenga conto delle varie tipologie di eventi che si conoscono.

2.2 Analisi strutturale di regioni globulari

Lo studio della struttura primaria fornisce un primo importante processo interpretativo per l'analisi funzionale. Ad essa si affiancano metodi di predizione di struttura secondaria e terziaria, che permettono di identificare substrutture date dal riarrangiamento di elementi di struttura secondaria (alfa eliche e beta strand), fino alla struttura dei domini. In particolare nel caso di pendrina è stata fatta un'analisi approfondita della regione transmembrana ampliando lo studio a quella che viene definita la dimensione 2,5 D. Vengono di seguito descritti solo gli strumenti computazionali che sono stati utilizzati in questo lavoro. Nel caso di proteine la cui struttura tridimensionale non è stata ottenuta sperimentalmente la predizione di struttura tridimensionale può essere applicata come metodo di fold recognition per identificare distanti relazioni evolutive tra strutture di domini proteici e funzioni, laddove PSI-BLAST non individua sequenze simili con struttura nota. Altrimenti le strutture secondarie e terziarie predette possono dare nuove ipotesi per le regioni di sequenza conservate o indicare potenziali siti di legame a proteine.

2.2.1 Metodi predittivi 1D

Data la sequenza di una proteina ci sono tutta una serie di predizioni che si possono fare che riguardano la struttura secondaria, l'accessibilità al solvente, il disordine, la presenza di repeats e la predizione di regioni transmembrana. Per l'analisi strutturale di Pendrina è stato condotto uno studio approfondito della regione transmembrana alla cui descrizione è stato dedicato un intero paragrafo.

Analisi di struttura secondaria

Predire la struttura secondaria di una proteina è il primo passo comunemente utilizzato per la sua classificazione ed il modelling. Per la predizione della struttura secondaria di Pendrina sono stati applicati PSIPRED e Porter (usato da SPRITZ) in modo da ottenere una predizione consensus.

PSIPRED è un metodo di terza generazione che utilizza i profili di PSI-BLAST ed è implementato con due reti neurali di tipo feed-forward. Calcola la propensione per struttura secondaria del residuo su una finestra di 15 aminoacidi. Il limite della lunghezza fissa dell'intorno locale rende difficile la predizione di foglietti beta.

Dal server **SPRITZ** è possibile ottenere la predizione di struttura secondaria e del disordine. La risposta viene fornita tramite email, sotto forma di testo organizzato secondo tre righe informative relative alla sequenza aminoacidica, la struttura secondaria e la predizione del disordine. Il protocollo **Porter** è un nuovo sistema per la predizione di struttura secondaria in tre classi: alpha-helix (H), extended (E), ovvero beta-strand e coil (C), che comprende tutto il resto. Porter si basa su reti neurali bidirezionali ricorrenti che permettono di allargare dinamicamente la finestra dell'intorno locale nel processo di valutazione alla propensione per una certa struttura secondaria. Questa caratteristica associata a un accurata codifica dei profili ottenuti da allineamenti multipli migliora la capacità di individuare significative relazioni tra residui distanti per strutture come i beta-strand.

Predizione del disordine

Lo stato disordinato è indicativo della presenza di regioni maggiormente plasmabili alle modificazioni indotte da variazioni ambientali o alle interazioni con ligandi. Anche in questo caso sono stati utilizzati due metodi diversi: SPRITZ e DISOPRED.

Il protocollo per la predizione del disordine di **SPRITZ** fornisce un'indicazione sulle regioni intrinsecamente disordinate (D) rispetto

a quelle ordinate (O) della proteina, mediante l'uso di classificatori binari implementati ad analisi probabilistiche mediante support vector machine.

Il protocollo **DISOPRED** è molto simile a quello di PSIPRED con il risultato della seconda rete a due stadi: disordinato e ordinato. L'analisi parte sempre da un profilo di PSI-BLAST con una prima rete neurale che valuta la propensione degli aminoacidi in una finestra di 15 residui in configurazione disordinata o ordinata. La seconda rete filtra i risultati della prima decretando la propensione a uno dei due stati. Disopred 2 estende il concetto di DISOPRED utilizzando una *support vector machine* e profili di sequenza.

Accessibilità al solvente

La superficie di accessibilità è determinata dalla superficie di van-der-Waals più la metà dello spazio che occupa una molecola d'acqua. Informazioni sulla superficie accessibile sono presenti nei file di **DSSP** in cui oltre alla struttura secondaria relativa a ogni residuo viene costruito un modello di ponti idrogeno, calcolati mediante la legge di Coulomb. La predizione di accessibilità della sequenza di Pendrina è stata ricavata dal server **ConSeq** che fornisce per ogni residuo la propensione allo stato exposed (e) o buried (b) secondo la predizione fatta da una rete neurale appositamente addestrata.

Una possibilità interessante per l'analisi delle proteine è quella di colorare la superficie accessibile con il potenziale elettrostatico. In questo modo si possono identificare possibili siti di interazione o individuare regioni cariche positivamente e negativamente. Pymol utilizza un plugin, **APBS tools**, per calcolare la superficie elettrostatica della struttura che si basa sull'equazione non lineare di Poisson-Boltzmann.

Ricerca di motivi Coiled-coil (CCM)

Segmenti coiled-coil sono α -eliche anfipatiche associate con le loro faccie idrofobiche. Sono formati da un pattern di 7 residui designati come **abcdefg** di cui la posizione **a** e **d** sono occupate da aminoacidi idrofobici, mentre le posizioni **e** e **g** da aminoacidi idrofilici.

Parry (1982), autore del primo metodo per la predizione di motivi coiled-coil, propose di pesare la propensione a formare coiled-coil di una sequenza in base alla corrispondenza con una PSSM derivata da questa distribuzione.

COILS estende questo approccio sostituendo la preferenza aminoacidica con la frequenza, introducendo una finestra di

scanning e una scala di punteggio contro i database di riferimento per ottenere una probabilità. Il metodo di predizione si basa sulla presenza di sequenze idrofiliche cariche per cui elevati valori di probabilità potrebbero essere dovuti all'elevata incidenza di residui carichi. La predizione è effettuata su finestre di 28, 21 e 14 residui. Le estremità di segmenti coiled-coil appaiono essere accuratamente identificate in finestre di 21 residui e in genere si assume che residui con probabilità maggiore del 50% fanno parte di segmenti coiled-coil. Dove possibile, sequenze relate alla proteina di interesse dovrebbero essere analizzate per la predizione di segmenti coiled-coil, utilizzando come input un allineamento multiplo. Attenzione: le sequenze devono essere relate nelle regioni con elevato score perché il confronto sia significativo.

Altri 3 programmi sono seguiti a COILS: PairCoil, MultiCoil e LearnCoil, costruiti aumentando la quantità di informazioni imputate, incorporando la correlazione tra coppie di residui nella *scoring matrix* allenata con un più grande set di dati coiled-coil. **MultiCoil** in aggiunta può differenziare coiled-coil di due o tre filamenti. Un ulteriore aumento di informazioni in input si è avuto con **Marcoil**, che calcola la probabilità da una HMM ed è correntemente il solo metodo senza finestra, che rimuove un limite dei precedenti programmi. Anche lo sviluppo di **PCOILS** vede un aumento di informazioni in input, sostituendo il confronto sequenza-profilo di COILS con un confronto profilo-profilo. Il profilo è derivato o da un MSA fornito in input o generato automaticamente con PSI-BLAST. Uno studio comparativo fra metodi di predizione coiled-coil, effettuato utilizzando un set più ampio di CCM corti e non solo quelli lunghi di tropomiosina, miosina e filamenti intermedi, dimostra che COILS e Multicoils forniscono risultati simili, mentre Multicoil è più restrittivo nell'assegnare probabilità. Marcoil e PCOIL superano gli altri e Marcoil fornisce i risultati migliori. Infine, **Paircoil 2**, versione più recente, ha risultati peggiori di Marcoil e PCOIL.

2.2.2 Banche dati di strutture proteiche

La prima banca dati di strutture di interesse biologico è la banca dati **PDB** (*Protein Data Bank*), attualmente situata presso la RCSB della Rutgers University, USA. In essa sono depositati *files* contenenti le coordinate atomiche determinate attraverso analisi cristallografiche ai raggi X o mediante applicazione di tecniche di spettroscopia NMR di proteine. Ogni file è identificato da un codice a 4 caratteri: un numero e tre caratteri alfa numerici. Il numero identifica la versione del file e i caratteri restanti sono scelti, dove possibile, in modo da ricordare i nomi delle strutture descritte nel file stesso. Alcuni files contengono più di una catena amminoacidica, sia perché possono contenere complessi di proteine sia perché

possono presentarsi come omopolimeri. In questi casi le diverse catene si distinguono attraverso uno specifico indicatore di catena (una lettera maiuscola). E' necessario a volte poter conoscere se si tratta di un omopolimero nella forma biologicamente attiva o se semplicemente i due monomeri sono il risultato della cristallizzazione e fanno parte di una stessa unità asimmetrica. Per questo è stato utilizzato il programma **PQS** (Protein Quaternary Structure) dell'EBI che considera parametri quali la superficie di contatto per discriminare contatti cristallografici da interazioni reali.

Altre banche dati di strutture sono state derivate da PDB per raccogliere le informazioni in modo mirato. **DSSP** per esempio, contiene informazioni sulle strutture secondarie di ogni entry del PDB. In particolare, per ogni file di coordinate del PDB, esiste un file DSSP con lo stesso codice.

Classificazione strutturale

Tra le banche dati derivate da PDB esistono in particolare tre grosse risorse che suddividono l'insieme delle strutture proteiche note classificandole secondo diversi livelli gerarchici in base alle loro caratteristiche funzionali, strutturali ed evolutive (SCOP, CATH e FSSP). Per la ricerca di modelli-templato nella predizione di struttura terziaria di Pendrina sono state consultate SCOP e CATH.

SCOP (Structural Classification of Protein), principalmente annotata a mano, classifica le proteine in 4 principali livelli: classe, Fold, Superfamiglia e Famiglie. La classe descrive il contenuto in struttura secondaria del dominio proteico. Il Fold raggruppa proteine con struttura secondaria paragonandole in riarrangiamento e connessione topologica. Proteine evolutivamente relate che mostrano forte relazione funzionale vengono classificate nella stessa superfamiglia anche se hanno una bassa identità di sequenza. Quando invece l'identità di sequenza è maggiore del 30% o nel caso funzione e struttura siano molto simili, le proteine vengono raggruppate nella stessa famiglia.

CATH (Class, Architecture, Topology and Homology) classifica le proteine in modo semiautomatico utilizzando SSAP (Sequential Structure Alignment Program), un programma per la ricerca di similarità strutturali basato sul confronto di direzione dei vettori C-beta-C-beta intraproteina. CATH classifica le proteine in diversi livelli gerarchici, di cui i quattro principali sono: Classe (C), Architettura (A), Topologia (T) e Superfamiglia di proteine Omologhe (H). In questo caso la Classe e la Architettura sono simili ai livelli di SCOP mentre la Topologia raggruppa cluster strutturali in base alle loro connessioni spaziali e al numero delle strutture secondarie. Le

Superfamiglie Omologhe sono costituite da proteine con similarità strutturale (gruppi di fold simili) e funzionale.

ModBase è un database di modelli tridimensionali di proteine ottenuti mediante un programma automatico, ModPipe, che utilizza PSI-BLAST e MODELLER per la costruzione del modello. Modbase include anche l'assegnazione del fold e gli allineamenti su cui si basano i modelli. I modelli non sono determinati sperimentalmente bensì calcolati teoricamente e contengono errori significativi, per questo occorre valutare attentamente la loro qualità. UniProt contiene nella entry della proteina ricercata anche il link per ModBase che indica le strutture da utilizzare come template per la costruzione del modello tridimensionale.

CATH permette di ottenere una lista di proteine con fold simile alla sequenza di interesse di cui è stata determinata sperimentalmente la struttura. Queste strutture sono state analizzate per individuare le regioni strutturalmente conservate e quelle variabili mediante sovrapposizione strutturale con **CE (Combinatorial Extension)** che cerca il migliore allineamento possibile tra due strutture mediante un approccio simile alle mappe di contatti ma a livello di otto residui.

2.2.3 Programmi di visualizzazione di struttura terziaria

Rasmol, **Jmol** e **Pymol** sono i programmi free maggiormente utilizzati per la visualizzazione grafica delle strutture tridimensionali delle proteine. Essi permettono di spaziare nell'analisi della molecola tramite ingrandimenti, visualizzazione di contatti, raffigurazione della superficie sulla base dei raggi molecolari atomici. Pymol è stato inoltre utilizzato per l'analisi della superficie elettrostatica del modello del dominio STAS di Pendrina. Le immagini proposte in questo lavoro provengono da Pymol, che garantisce una grafica ad alta qualità.

2.2.4 Predizione 3D

Per identificare proteine template che appartengono alla stessa classe di fold della sequenza di interesse è stata fatta una prima ricerca con il **Meta-Server BioInfo.PL**, utilizzando come input la sequenza dei domini e non l'intera sequenza della proteina. Il server utilizza la predizione di diversi programmi e il modello predetto è valutato con il sistema **3D-Jury**, basato su un sofisticato schema di punteggio. L'obiettivo di programmi di *fold recognition* è quello di identificare proteine con strutture simili anche se la loro similarità di sequenza è bassa. La sequenza target infatti è confrontata con le informazioni provenienti dalla struttura tridimensionale nota. La

qualità delle strutture sperimentali identificate è stata valutata con TAP. Questo programma valuta per ciascun aminoacido la corrispondenza tra i valori degli angoli torsionali della struttura data in input con i valori degli angoli che sono stati più comunemente riscontrati in tutte le proteine note.

Una volta identificato il template sono stati utilizzati **ALIGN**, **SSEA** e **GenAltAli** per l'allineamento target – template. Questi programmi utilizzano le informazioni di struttura secondaria di entrambe le sequenze. La struttura secondaria del template viene ricavata da DSSP. **ALIGN** è stato utilizzato in modalità di allineamento profilo-profilo e algoritmo globale. Nell'allineamento ottenuto con **SSEA** gli elementi di struttura secondaria sono mantenuti a blocchi. Il vantaggio è che nel modello non ci saranno buchi a livello di alpha-eliche o beta-strand. **GenAltAli** opera in modo iterativo cambiando ogni volta i parametri di penalità per gap ed estensione entro un range prestabilito e dopo 625 cicli restituisce una lista di probabili allineamenti target – template in cui le regioni che presentano maggiore variabilità vengono escluse dall'allineamento. Dai diversi allineamenti ottenuti vengono costruiti automaticamente dei modelli grezzi con **Ali2Eval** che utilizza FRST come potenziale statistico per la valutazione del modello. Questo approccio è stato dimostrato migliorare mediamente i risultati del modelling (Sommer et al., 2006). Il migliore allineamento è stato poi aggiustato a mano in base al contesto sequenza-struttura, rispettando la geometria dei residui conservati e considerando la struttura secondaria. Inserzioni o delezioni possono essere spostate in modo da non interrompere o se necessario estendere le alpha-eliche e i beta-strand.

A questo punto la predizione del ripiegamento tridimensionale della sequenza aminoacidica viene eseguita mediante **HOMER** (*Homology Modelling Server*) sviluppato dall'Università di Padova. Al programma viene dato in input il migliore allineamento target-template e il codice PDB della struttura che si vuole usare come template che il programma utilizza per la costruzione del modello 3D del target. Il modello ottenuto può essere direttamente visualizzato con Jmol oppure è possibile scaricare il file pdb e visualizzarlo con Pymol.

2.3 Analisi strutturale di regioni transmembrana

La struttura di proteine transmembrana risulta difficile da determinare sperimentalmente con le note tecniche di cristallizzazione a Raggi X e NMR. Infatti soltanto l'1% delle strutture proteiche tridimensionali depositate in Protein Data Bank è relativo

a proteine transmembrana. Questo rappresenta un forte stimolo per lo sviluppo di efficienti metodi di predizione strutturale, semplificato dal fatto che proteine integrali si presentano principalmente in sole due tipologie di Fold: fascio di α -eliche e β -barrell. Entrambi i tipi inoltre mostrano delle caratteristiche simili nella composizione amminoacidica che deve rispettare l'esigenza di abbassare il costo energetico dovuto all'inserzione della proteina in un ambiente idrofobico.

2.3.1 Composizione amminoacidica

Dalle prime strutture ottenute con metodi ad alta risoluzione, per le proteine di membrana a fascio di α -eliche, si sono immediatamente individuate alcune caratteristiche peculiari nella composizione amminoacidica delle diverse regioni strutturali. I tipici segmenti che attraversano la membrana sono costituiti prevalentemente da sequenze di residui idrofobici. Gli amminoacidi aromatici (**Trp e Tyr**) tendono a clusterizzare alle estremità dei segmenti transmembrana; i loop che connettono le eliche differiscono nella composizione amminoacidica in base al lato della cellula su cui si affacciano. Il termine **positive-inside rule** è stato coniato appunto per descrivere la particolare concentrazione di amminoacidi carichi positivamente nei loop intracellulari.

2.3.2 Predizione della topologia

I primi metodi di predizione usavano un semplice plot di idrofobicità per individuare segmenti transmembrana con una finestra fissa di 19 residui (**Kyte and Doolittle method**). Ad oggi soprattutto per proteine a fascio di α -eliche sono stati sviluppati metodi in grado di predire la topologia di membrana. Questo consiste nel determinare non solo i segmenti che attraversano la membrana, ma anche l'orientamento dei loop rispetto all'interno-esterno della cellula. Per la predizione della topologia di membrana della Pendrina sono stati impiegati alcuni dei più evoluti metodi computazionali qui di seguito descritti.

TopPred (von Heijne, 1992) combina le informazioni derivanti dal plot di idrofobicità con la positive-inside rule secondo cui i loop intracellulari mostrano una maggiore percentuale di Lys e Arg rispetto ai loop esterni. Permette di scegliere la topologia migliore tra poche alternative.

TMpred utilizza un algoritmo basato sull'analisi statistica di TMbase, un database di proteine transmembrana, e combina diverse matrici di peso per assegnare il punteggio al modello migliore.

MEMSAT estende il concetto di propensione degli amminoacidi a regioni della membrana e calcola per ogni segmento transmembrana lunghezza, localizzazione e orientamento topologico più probabili, utilizzando un algoritmo di programmazione dinamica che massimizza il punteggio per ogni singolo amminoacido secondo le propensioni a 5 classi distinte. Per l'analisi di Pendrina è stata utilizzata l'ultima versione **MEMSAT 3** (Jones, 2007) che aggiunge informazioni evolutive derivanti dai profili di sequenza ottenuti con PSI-BLAST su cui viene allenata una rete neurale in grado di predire i segmenti transmembrana, assegnare il punteggio ai modelli topologici creati e identificare possibili peptidi segnale.

TMHMM (Krogh, 2001) e **HMMTOP** sono i primi metodi che impiegano le Hidden Markov Model (HMM) come metodo di machine learning per dedurre delle regole nella struttura delle proteine transmembrana. Il vantaggio delle HMM è la possibilità di modulare la lunghezza dei segmenti transmembrana e incorporare informazioni relative alla composizione amminoacidica. A seconda della zona in cui si trovano blocchi di amminoacidi adiacenti mostrano caratteristiche diverse e le HMM vengono allenate sulla base di queste distinte propensioni. TMHMM implementa un modello ciclico con 7 stati per eliche transmembrana, mentre HMMTOP usa HMM per distinguere tra 5 stati strutturali [helix core (H), inside loop (i), outside loop (o), helix caps (C e N) e domini globulari (O)]. Per la predizione della topologia di Pendrina sono state utilizzate anche le ultime versioni, **TMHMM 2.0** e **HMMTOP 2.0** (Tusnàdi, 2001). Quest'ultima permette di imporre la posizione di un amminoacido la cui localizzazione è stata sperimentalmente determinata.

Phobius (Kall, 2004) è un altro metodo basato su HMM che però è in grado di distinguere tra sequenze segnale e segmenti transmembrana e come HMMTOP permette di aggiungere informazioni sperimentali alla predizione (**Phobius constrained**).

Tra i metodi basati su HMM sono stati recentemente sviluppate versioni che includono informazioni evolutive come **Prodiv TMHMM** e **PoliPhobius** (Kall, 2005) che implementa il metodo con informazioni evolutive derivanti da un multiallineamento di sequenze omologhe ricercate in UNIProt/TrEMBL.

L'assegnazione dei limiti di segmenti transmembrana è stata infine condotta mediante il metodo consenso che permette di porre come più probabili i limiti predetti dal maggior numero di metodi utilizzati. A tal fine sono stati considerati anche alcuni server (**Sfinx** e **Pongo**) che mettono a disposizione il confronto grafico dei risultati di predizione di topologia di almeno cinque metodi predefiniti.

I segmenti transmembrana di dubbia predizione sono stati valutati utilizzando l'editor Jalview per l'allineamento con la sequenza proteica delle informazioni ottenute con i metodi di predizione di struttura secondaria, del disordine, dell'accessibilità al solvente e delle regioni transmembrana, modificando laddove necessario l'output ottenuto dai diversi metodi.

2.3.3 Predizione 2,5D

Dalle strutture di proteine transmembrana ottenute sperimentalmente è stato osservato che alcuni segmenti attraversano la membrana con un percorso diverso rispetto all'asse perpendicolare del doppio strato lipidico o non lo attraversano completamente. In particolare sono state individuate regioni rientranti (RR) ed eliche interfaccia che in diversi casi rappresentano elementi strutturali essenziali per la funzione della proteina. Si pensa che le RR si formino relativamente più tardi nella dinamica del folding. Dopo l'iniziale traslocazione e formazione delle eliche transmembrana, una parte dei loop viene spinta nella membrana dove interagiscono con le eliche transmembrana (TM) altamente idrofobiche per minimizzare la penalità energetica imposta dalla presenza di residui polari o carichi in un ambiente a basso indice dielettrico (Lasso et al. 2006). La maggior parte delle RR infatti è situata nelle regioni dei loop più corti di 50 residui, tra una TM e l'altra. Nelle RR si riscontra la presenza preferenziale di residui idrofobici e una ridotta propensione per amminoacidi carichi, caratteristiche simili a quelle delle TM che rendono difficile la distinzione tra i due tipi di strutture, ma frequentemente le RR sono anfifiliche e contengono sezioni idrofobiche troppo intermittenti per essere identificate come regioni di membrana (Lasso et al., 2006). Caratteristica peculiare delle RR è la presenza di molti amminoacidi di piccole dimensioni, come alanina e glicina, soprattutto nelle regioni coil, mentre un elevato contenuto di Tyr e Trp è una caratteristica delle regioni interfaccia. Queste strutture si dispongono parallelamente all'asse della membrana a livello dell'interfaccia tra la parte acquosa intra o extracellulare e il doppio strato lipidico. I residui aromatici si dispongono con la parte polare verso le teste polari dei fosfolipidi e la parte idrofobica immersa nello strato lipidico.

La frazione di elementi di struttura secondaria irregolare è maggiore nelle regioni rientranti rispetto alle regolari eliche transmembrana. Comunque, in base alla struttura secondaria si possono distinguere 3 classi di regioni rientranti: lunghe regioni con elica-coil-elica (H-C-H); regioni di media lunghezza con motivi elica-coil (H-C) o coil-elica (C-H); Brevi regioni con struttura secondaria interamente

irregolare. Le RR H-C, C-H e H-C-H sono più facilmente rilevabili di quelle con struttura secondaria irregolare, in quanto possono essere identificate erroneamente come TM a causa del pattern di idrofobicità simile. Le RR lunghe con struttura HCH, più corte di un paio di TM, possono invece essere interpretate come un'unica TM. E' possibile però aumentare la probabilità di identificare una RR se è nota la posizione delle TM (Viklund et al., 2006)

2.3.4 Predizione di Regioni rientranti

Nello studio della regione transmembrana di Pendrina è stato quindi incluso l'utilizzo dei nuovi metodi computazionali che sono in grado di distinguere tra segmenti transmembrana e le regioni rientranti.

TMLOOP (Lasso et al., 2006) utilizza un approccio di *pattern discovery* per identificare le regioni rientranti che viene applicato ad un set di famiglie proteiche contenenti regioni rientranti con struttura nota. Il pattern scoperto sembra essere composto di residui il cui ruolo biochimico è noto essere essenziale per la funzione della proteina, e questo non sorprende in quanto residui conservati sono spesso indicativi di un comune ruolo strutturale o funzionale. L'identificazione del pattern viene condotta con tre diverse strategie: i) *exact pattern discovery* ii) *pattern discovery using a chemical equivalence set* iii) *pattern discovery using a chemical structural set*. I patterns identificati vengono usati come regola di predizione per predire regioni rientranti in proteine politopiche, ovvero con più segmenti transmembrana, con un'accuratezza del 92,4% e 100% di affidabilità. Come input sono state inserite 100 sequenze omologhe alla Pendrina e sono stati utilizzati i due tipi di approccio che mette a disposizione il programma: il metodo a singolo motivo, che cerca l'esatta sovrapposizione al pattern, e il metodo collettivo, basato sull'uso di diversi pattern, che consente di individuare anche regioni che presentano piccole variazioni rispetto al pattern singolo.

ZPRED (Granseth et al., 2006) per la predizione di Z-coordinate, calcola la distanza per ciascun residuo dal centro della membrana ma non la sua direzione. Il metodo assume alcune semplificazioni: residui che sono a 0-5Å si trovano nella regione idrofobica centrale mentre quelli che superano i 25Å si trovano in un ambiente non di membrana. Questo significa che il predittore si focalizza sulle regioni interne alla membrana in cui l'ambiente presenta le maggiori variazioni. ZPRED usa l'output di una HMM come input per una rete neurale assieme a il profilo di informazioni di una finestra di sequenza, ottenendo un errore medio di 2.55Å. La localizzazione delle TM viene predetta con Prodiv-TMHMM.

TOPMOD (Viklund et al., 2006) ha come obiettivo individuare le RR sulla base delle sole caratteristiche strutturali per cui divide ogni sequenza in 4 regioni strutturali [membrana (M), rientrante (R), elica interfaccia (I) e loop (L)] e utilizza una HMM per classificare i residui di una sequenza proteica nelle quattro classi strutturali e per predire le RR. Utilizza pochi parametri liberi per evitare l'overfitting data la scarsità di dati. Tre componenti sono sufficienti a descrivere completamente le differenze tra i quattro stati: una componente che riflette le differenze in idrofobicità, una componente che discrimina fra composizione in amminoacidi prevalentemente piccoli o grandi (valori molto positivi per Gly, Ala, Thr e Asn; valori molto negativi per Leu, Trp, Ile e Phe); la terza componente individua le differenze nel contenuto di residui aromatici (Tyr e Trp) caratteristica per le regioni interfaccia. Anche TOPMOD considera le regioni transmembrana note a priori e utilizza ProdivTMHMM per la predizione.

2.3.5 Predizione dell'orientamento delle eliche nella membrana

In contrasto alle proteine globulari, le proteine transmembrana non mostrano molta differenza tra residui esposti ai lipidi e quelli nascosti nelle regioni impaccate. Così risulta arduo prevedere la superficie di accessibilità. Le caratteristiche che permettono maggiormente di distinguere tra residui esposti ai lipidi e quelli non, sono la polarità delle catene laterali e il grado di conservazione della sequenza. Residui maggiormente idrofobici e meno conservati tendono ad essere più esposti ai lipidi (Elofsson et al., 2007).

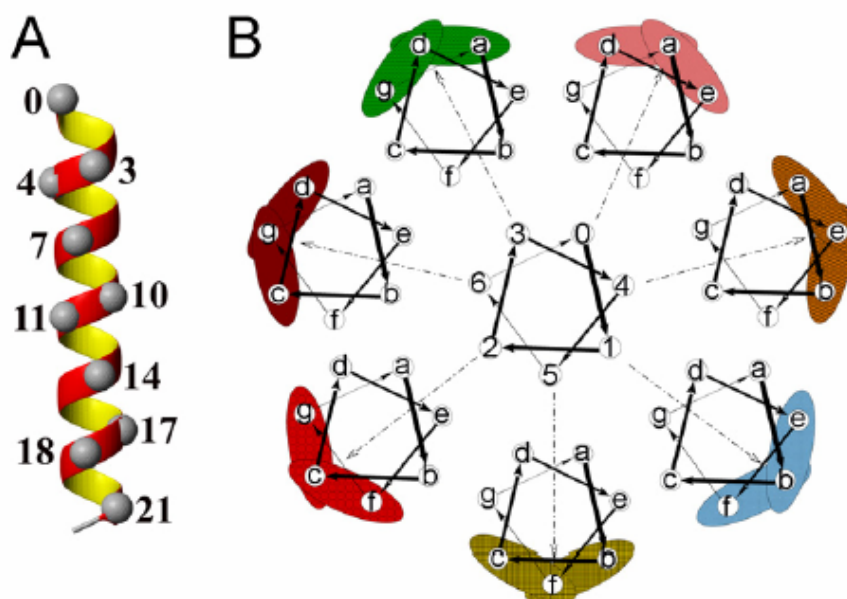


Fig. 2.1 Rappresentazione schematica delle sette facce che compongono le eliche transmembrana secondo il metodo LIPS. (Adamian et al., 2006)

LIPS (Adamian et al., 2006) è un metodo per la predizione dell'orientamento delle eliche TM basato su un modello canonico a sette facce sviluppato originariamente per strutture coiled coil (Fig. 2.1).

Ad ogni faccia viene assegnato un punteggio calcolato dalla scoring function LIPS (LIPid-face in surface) che combina una scala di lipofilicità (TMLIP2) e conservazione dei residui nell'elica. L'orientamento dell'elica viene così predetto cercando la faccia più esposta ai lipidi o quella più nascosta.

Il programma richiede in input un multiallineamento (MSA) della regione corrispondente ad un'elica transmembrana. Per la costruzione del MSA di sequenze omologhe di Pendrina è stata fatta una ricerca con PSI-BLAST nel database di sequenze non-ridondanti con 4 iterazioni. Da questo primo set di sequenze sono state selezionate e allineate con ClustalW 100 sequenze con identità inferiore a 95%. Per la scelta delle eliche TM mi sono basata sui risultati ottenuti dal metodo consenso per la predizione della topologia, utilizzando Jalview come editor per produrre blocchi separati di sequenze privi di gap per ciascuna elica TM. L'output riporta il punteggio di LIPS per tutte le 7 possibili facce.

Il server **RANTS** (Adamian et al., 2006) permette di classificare le eliche di proteine politopiche secondo il grado di accessibilità al solvente. In questo caso il punteggio maggiore va all'elica meno esposta ai lipidi. Il programma restituisce in output un valore per ogni elica con cui ho costruito un grafico per la valutazione dei dati.

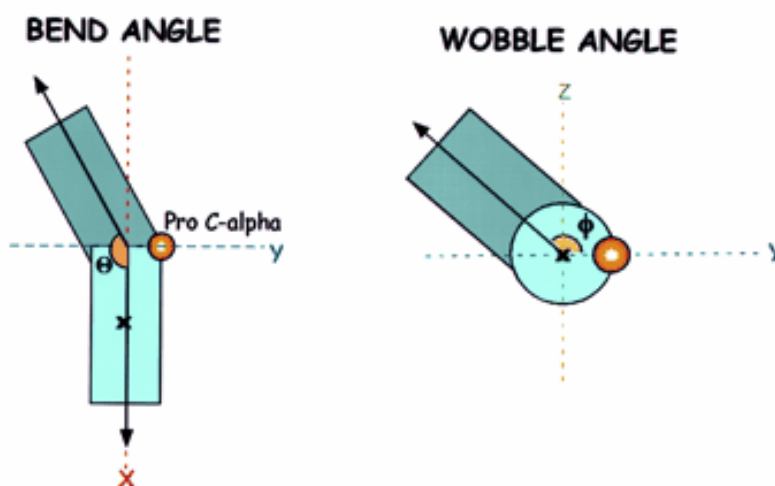


Fig. 2.2 Distorsione indotta da prolina in un'elica di membrana. (Visiers et al., 2000)

2.3.6 Presenza di proline-induced-kinks

La presenza nelle eliche TM di un'incurvatura rappresenta un'ulteriore caratteristica 2,5D che può essere predetta con ragionevole accuratezza. Tali incurvature possono essere indotte dalla presenza di una prolina, la cui catena laterale preclude la normale geometria dell'elica, e si conservano anche quando la prolina viene sostituita con un altro residuo, come l'alanina. La presenza di una prolina in una particolare posizione in almeno il 10% delle sequenze di un multiallineamento permette di predire con una certa affidabilità/confidenza che si tratti di un sito di curvatura. Si è ipotizzato che nel corso dell'evoluzione, la curvatura inizialmente avvenga per sostituzione di un residuo con una prolina. La distorsione dell'elica che ne deriva (Fig. 2.2), in certe posizioni, può essere tollerata o assumere un valore adattativo che viene conservato nelle generazioni future (Yohannan et al., 2003).

2.3.7 Presenza di motivi GxxxG

Recentemente sono stati individuati motivi di sequenza, come il motivo **GxxxG**, e pattern periodici, statisticamente frequenti nelle proteine integrali, che sembrano essere coinvolti nelle interazioni fra eliche e nella formazione della struttura tridimensionale. La glicina è nota per fungere da cerniera flessibile da cui viene distrutta la struttura ad alfa elica di un peptide posto in ambiente acquoso ed è presente in diversi motivi di sequenza che mediano la dimerizzazione. Si possono trovare anche più motivi adiacenti, come GxxxGxxxG dove le glicine esterne possono essere sostituite da altri residui purché di piccole dimensioni (Ala, Ser, Thr). Il motivo GxxxG posiziona due glicine sullo stesso lato dell'elica (Fig. 2.3) o sulla stessa faccia del β -sheet. In assenza di ingombro della catena laterale si ha un contatto in entrambi i tipi di struttura secondaria.

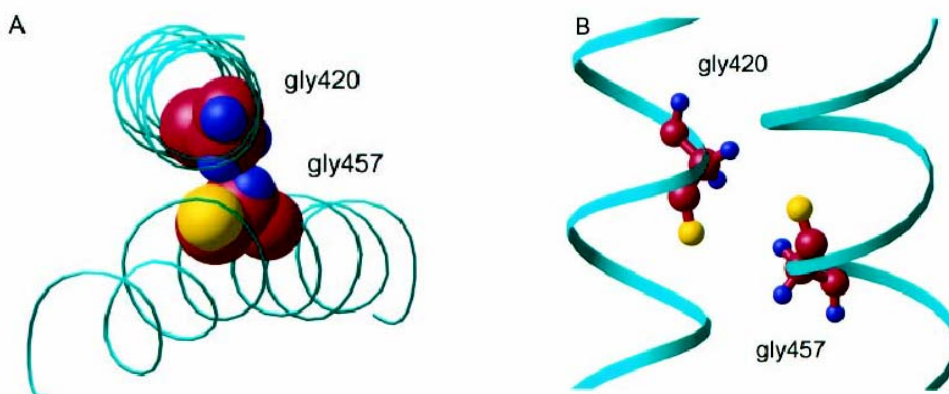


Fig. 2.3 Esempi di interazione tra motivi GxxxG in eliche transmembrana (Javadpour et al., 199)).

Motivi GxxxG sembrano essere responsabili della formazione di fibrille amiloidogeniche, strutture che si formano per conversione di α -eliche transmembrana in foglietti- β , come accade nella Sindrome di Alzheimer (Liu et al., 2005).

Allo scopo di individuare putativi siti di interazione elica-elica e siti di probabile dimerizzazione nella regione transmembrana di Pendrina, i motivi GXXXG e quelli degenerati sono stati mappati e valutato il loro ruolo in relazione alle informazioni ottenute dai metodi di predizione utilizzati.

2.4 Analisi dell'architettura funzionale delle proteine

Recentemente si assiste ad un crescente sviluppo di metodi computazionali sempre più sofisticati capaci di esplorare la funzione di una proteina che deve essere pensata su differenti livelli interdipendenti e può essere suddivisa in tre principali categorie: funzione molecolare, processo biologico e componente cellulare. La funzione molecolare descrive l'attività della proteina a livello molecolare. L'assemblamento di diverse funzioni molecolari determina il processo biologico in cui essa è coinvolta mentre la componente cellulare individua il compartimento della cellula in cui la proteina svolge la sua funzione.

Metodi ad inferenza genomica possono identificare direttamente le interazioni proteina-proteina o indirettamente le associazioni funzionali che sono state trovate in processi biologici. La componente cellulare può essere individuata mediante metodi per la predizione di sequenze segnale, la composizione in amminoacidi, l'associazione alla membrana o modificazioni post-traduzionali.

Un approccio comunemente utilizzato nella predizione di funzione si basa sul concetto che proteine con sequenze simili molto spesso mostrano funzioni simili. Per l'identificazione di ortologhi si possono consultare risorse di annotazioni funzionali, come GO (Gene Ontology) e risorse che raggruppano sequenze di domini in famiglie evolutivamente relate.

Quando la funzione è dedotta dalla similarità strutturale la questione è più complessa. Esistono proteine che mostrano lo stesso fold e sito attivo ma hanno funzioni differenti e dal lato opposto le medesime funzioni trovano differenti soluzioni strutturali, dove non esiste un comune progenitore si ha una evoluzione convergente.

Nonostante ciò, metodi per la sovrapposizione strutturale in combinazione con la ricerca di motivi strutturali e di siti funzionali risultano cruciali per identificare relazioni evolutive dove la similarità di sequenza non è significativa.

2.4.1 Architettura dei domini della proteina

Un comune approccio per predire la funzione di una data sequenza proteica è usare i gateway forniti da NCBI ed EMBL-EBI. Queste risorse guidano l'operatore alla raccolta di dati accurati, incluse informazioni sulle famiglie di domini e di proteine, siti funzionali e metodi di predizione della funzione. Attraverso il software **InterPROscan** è possibile ricercare motivi strutturali e funzionali annotati nelle banche dati, integrate in Interpro, che raggruppano sequenze multi-domini o domini proteici individuali in putative famiglie evolutive. Tra le diverse risorse disponibili, Pfam, SCOP, SUPERFAMILY e CATH sono state consultate per lo studio funzionale di Pendrina. SCOP e CATH sono stati precedentemente descritti.

Pfam definisce famiglie di domini proteici basandosi su hidden Markov Models (HMMs) derivate da allineamenti multipli di sequenze. Pfam è contenuto anche nel database di Domini Conservati (**CDD**) di NCBI ma in questo caso la ricerca di domini avviene mediante matrici di punteggio posizione-specifica (PSSMs) derivati da Pfam e SMART. I risultati di CDD sono stati ottenuti mediante il server PAT per la predizione 1D di sequenze proteiche.

Le banche dati di strutture proteiche sopra descritte sono state consultate per la ricerca di proteine con *fold* simile allo scopo di individuare le differenze strutturali responsabili dell'evoluzione di differenti funzioni. E' noto infatti che in superfamiglie di proteine altamente variabili, differenti funzioni possono evolvere grazie all'inserzione di elementi di struttura secondaria che nella struttura tendono a collocarsi in specifici punti per produrre un dominio più grande o una particolare superficie che modifica la geometria del sito attivo e promuove differenti interazioni proteina-proteina.

CATH è stato usato inoltre per individuare altre proteine contenenti il dominio STAS al fine di identificare domini proteici coevoluti che originariamente facevano parte di uno stesso pathway metabolico.

2.4.2 Identificazione di motivi lineari

Proteine multi dominio di grandi dimensioni si possono presentare disordinate prive di una conformazione regolare. A volte il disordine è presente semplicemente nelle regioni che connettono domini globulari in cui spesso sono presenti siti funzionali come siti di

interazione tra proteine, segnali di localizzazione cellulare, siti di modificazione post-traduzionale o siti di taglio. Questi piccoli motivi vengono descritti e identificati mediante pattern e espressioni regolari e sono importanti perchè possono suggerire delle ipotesi funzionali. Un problema generale, riguarda l'elevata incidenza di falsi positivi dovuta alla bassa significatività statistica di corte sequenze che ricorrono frequentemente. Per la ricerca di pattern funzionali di sequenza, di interazioni con ligandi o per eventuali modifiche post-traduzionali vengono generalmente svolte mediante i database **PROSITE** o **ELM**. In particolare è stato utilizzato anche **Phospho-ELM**, un sistema di ricerca specifico per siti di fosforilazione. La presenza di siti di glicosilazione è stata valutata con **Net N-glyc**, per glicosilazioni che interessano l'asparagina, e **Net O-glyc**, per glicosilazioni serina o treonina. Per l'analisi della localizzazione cellulare viene utilizzato **PSORT II** che è in grado di predire potenziali peptidi segnale per la ritenzione entro il reticolo endoplasmatico o vescicolare e segnali per la localizzazione nucleare. **SignalP** e **TargetP** accompagnano la predizione di sequenze segnale caratterizzate dalla presenza di siti di taglio post-traduzionali. A livello della sequenza dello STAS domain di Pendrina è stato inoltre ricercato manualmente il motivo di legame al colesterolo $[L/V]-x(1-5)-Y-x(1-15)-[R/K]$.

2.4.3 Studio di residui discriminanti per la funzione

Il server **ConSeq** costruisce un multiallineamento di proteine con Clustalw e associa ad ogni residuo una gradazione di colore in base al grado di conservazione. Questo programma è stato utilizzato nella definizione dei segmenti transmembrana e per avere un riscontro immediato della conservazione di potenziali motivi individuati con i diversi predittori. Il server **Consurf** invece visualizza la conservazione aminoacidica direttamente nel modello, secondo la stessa scala di colorazione. Per l'analisi del modello del dominio STAS di Pendrina è stato dato in input un multiallineamento curato a mano di sequenze omologhe di Pendrina e il file PDB del modello proposto. La visualizzazione viene fornita direttamente dal server mediante Jmol. L'elaborazione di Consurf è stata utilizzata nell'analisi funzionale del dominio STAS. Aminoacidi che si trovano in regioni conservate hanno un potenziale ruolo strutturale e funzionale per cui tendono a essere mantenuti tali durante l'evoluzione. In particolare i residui maggiormente conservati che vengono a trovarsi vicini nello spazio tridimensionale possono essere indicativi di un sito funzionale.

3. Risultati

La sequenza aminoacidica di Pendrina è stata ottenuta mediante l'interrogazione della banca dati GenBank di NCBI. L'analisi con PSIBLAST ha permesso di raccogliere sequenze di proteine omologhe a Pendrina con cui, mediante il programma MAFFT, è stato costruito un multiallineamento (MSA) utilizzato per le successive fasi di analisi strutturale e per costruire l'albero filogenetico della famiglia dei trasportatori del solfato SLC26. Attraverso il software InterProScan è stato possibile ricercare i domini strutturali e funzionali, annotati nelle banche dati integrate in Interpro, e stabilire esattamente la loro localizzazione sulla sequenza di Pendrina. Si è deciso di analizzare le informazioni ottenute dall'applicazione dei diversi metodi bioinformatici precedentemente descritti suddividendo la proteina in tre distinte regioni: estremità N-terminale, regione transmembrana e l'estremità C-terminale. Queste porzioni in Pendrina mostrano differenti caratteristiche e contengono al loro interno domini distinti. E' stato quindi possibile applicare per ciascuna regione una adeguata strategia di analisi. La sequenza di Pendrina è stata analizzata per la predizione di struttura secondaria, del disordine e di strutture coiled-coil. Per la regione transmembrana è stata condotta un'analisi approfondita per la predizione della topologia di membrana che ha permesso di stabilire i limiti di ciascun segmento transmembrana. Sono stati inoltre utilizzati programmi per la predizione di regioni rientranti ed eliche interfaccia e la sequenza di Pendrina è stata valutata rispetto al contenuto aminoacidico per individuare eventuali caratteristiche tipiche delle regioni rientranti o di Proline all'interno delle eliche di membrana che potrebbero dare un'indicazione della struttura 2,5D di queste regioni. Le sequenze di ciascuna elica sono state poi caratterizzate, mediante LIPS, per identificare l'orientamento delle eliche rispetto all'ambiente idrofobico della membrana e mediante RANTS per classificare le eliche in base alla loro accessibilità ai lipidi. Sulla sequenza di Pendrina infine sono stati mappati motivi GXXXG e per quelli localizzati nelle eliche interfaccia predette è stata valutata la disposizione dei residui di Glicina rispetto alle facce potenzialmente coinvolte nelle interazioni con altre eliche.

Per il dominio transmembrana e per il dominio STAS è stata eseguita una ricerca con Metaserver al fine di identificare un fold che presenti una struttura secondaria il più affine possibile a quella trovata per Pendrina nei diversi domini. Solo nel caso del dominio STAS il server ha indicato un punteggio elevato per alcune strutture della famiglia di *fold* SpollAA-like. In base alla qualità della struttura

e all'identità di sequenza con Pendrina, è stato scelto il modello della proteina SpollAA (codice pdb 1h4y), un antagonista del fattore anti-sigma di *Bacillus Sphaericus*, come template per la costruzione del modello tridimensionale del dominio STAS. L'allineamento delle sequenze di SpollAA e del dominio STAS, ottenuto con GenAltAli e curato manualmente, è stato utilizzato per la realizzazione del modello tridimensionale con HOMER-M. Infine il modello è stato perfezionato mediante minimizzazione energetica e caratterizzato tramite il server Consurf per la predizione della conservazione aminoacidica.

Una volta completata l'analisi strutturale, sono state mappate le mutazioni di Pendrina trovate in soggetti con Sindrome di Pendred e sordità di tipo DFNB4. Un'analisi approfondita è stata proposta per le mutazioni che interessano il dominio STAS. Utilizzando il modello tridimensionale predetto è stato possibile ipotizzare i possibili effetti della sostituzione aminoacidica sulla struttura e funzione di Pendrina.

3.1 Pendrina e la filogenesi della famiglia dei trasportatori SLC26A.

Pendrina è costituita da 780 aminoacidi con un dominio transmembrana centrale e le estremità N e C terminali intracellulari. La ricerca nelle banche dati di domini tramite InterProScan ha permesso di identificare la presenza dei domini e motivi conservati presenti nella sequenza di Pendrina. La banca dati TIGRFAM riconosce un unico grande dominio, il dominio della famiglia SLP (*TIGR00815*), che comprende la regione transmembrana e parte dell'estremità C-terminale. Pfam invece distingue il dominio dei trasportatori del solfato nella parte C-terminale della regione transmembrana (*PF00916*) e il dominio STAS (*PF01740*), all'estremità C-terminale. Superfamily Sequence Search individua un'unica corrispondenza tra la sequenza dell'estremità C-terminale di Pendrina e il dominio STAS (*SSF5209*). PROSITE identifica il motivo dei trasportatori del solfato SLC26A (*PS01130*) e il dominio STAS (*PS50801*). Un allineamento multiplo di Pendrina e altri membri SLC26A è rappresentato in Fig. 3.1.

Filogenesi

L'albero ottenuto applicando il metodo di Maximun Likelihood (ML) al multiallineamento di sequenze del dominio STAS dei trasportatori del solfato SLC26A è presentato in Fig. 3.2. Dall'albero ottenuto si osserva che le sequenze del dominio STAS di batteri assumono una

S26A4_HUMAN	1	10	20	30	40
S26A4_HUMAN	MAA.....PGRSEPPQLPEYSCSYMVSRPVYSBLAFQQ.....QHRRRLQBRK				
S26A4_RAT	MAA.....RDRSEPPQLAEYSCSYAVSRPVYSBLAFQQ.....QRRRLPERR				
S26A4_MOUSE	MAA.....RGRSEPPQLAEYSCSYTVSRPVYSBLAFQQ.....QRRRLPERR				
S26A5_HUMAN	MDH.....ABENEILAAATQRYVVERPIFSHPVLQE.....RLHTKDKVPD				
S26A3_HUMAN	MLE.....PPGNQYIVARPVYSTNAFEE.....NHKKTGRHHK				
S26A6_HUMAN	MGLADASGPRDTQALLSATQAMDRLRRRDYHMERPLLNQEHLEE.....LGRWGSAPR				
S26A2_HUMAN	MSSESEQHNVSPRDSABGNDYS..PSGIHLELQRESSTDFKQFETNDQCRPYHRIIERQEKSDTNFKB				
S26A1_HUMAN	MDE.....SPEPLQQGRGPV.....PVRQRPAFRGLRE				

S26A4_HUMAN	50	60	70	80	90	100	110
S26A4_HUMAN	TLRESLAKCCS	CSRKRAFGVLKTLV	FILEWLPK	RVKEWLLS	DIISGVSTGLVATL	QGMAYAL	LAAPVVG
S26A4_RAT	TLRDSLARS	CSRKRAFGALKALL	PILDWLPK	RVKEWLLS	DIISGVSTGLVGT	QGMAYAL	LAAPVVG
S26A4_MOUSE	TLRDSLARS	CSRKRAFGVVKTL	PILDWLPK	RVKEWLLS	DIISGVSTGLVGT	QGMAYAL	LAAPVVG
S26A5_HUMAN	SIADKLQAPT	CTPKKIRNIYMP	ITKWLPA	KFKKEYVLG	DLVSGISTGVQLP	QGLAFAL	LAAPVPI
S26A3_HUMAN	TFDLHLKVC	CSPOKAKRIVLS	PIASWLP	RLKEWLLS	DIISGVSTGVAVL	QGLAFAL	LVDIPPV
S26A6_HUMAN	TQWR..TWLQ	CSRARAYALLQ	HLPLVWLP	PRVPRDWLL	QGLSGLSVAINQ	PLQGLAYAL	LAGLPV
S26A2_HUMAN	FVIKKLQKNC	QSPAKAKNMIL	GFLPVLOWL	PKYDLKKNIL	GDMVSGLVIGIL	LVPSIAYSL	LAGQBPV
S26A1_HUMAN	MLKARLWC	CSVLQCVRALV	QDILLPATRWLR	QYRPREYLAG	DVMGSLVIGIL	LVPSIAYSL	LAGLQPI

S26A4_HUMAN	120	130	140	150	160				
S26A4_HUMAN	YGLYSAFP	PILTYFIF	GTSRHS	ISVGF	FPVVS	LMVGS	SVVLSMA	PDEHFLV	SSSN.....
S26A4_RAT	YGLYSAFP	PILTYFVF	GTSRHS	ISVGF	FPVVS	LMVGS	SVVLSMA	PDDHFLV	PSGN.....
S26A4_MOUSE	FGLYSAFP	PILTYFVF	GTSRHS	ISVGF	FPVVS	LMVGS	SVVLSMA	PDDHFLV	PSGN.....
S26A5_HUMAN	FGLYSFP	YIMYCFL	GTSRHS	ISVGF	FAVIS	LMIG	SVAVRLV	PDDIVIP	GGVN.....
S26A3_HUMAN	YGLYSFP	PAIYLF	GTSRHS	ISVGF	FPILS	MMVGL	AVSGAVS	KAVPDRN	ATT.....
S26A6_HUMAN	FGLYSFP	YFVIF	FLGTSRHS	ISVGT	FAVMS	VMVGS	VTESLA	PQALNDS	MINR.....
S26A2_HUMAN	YGLYTS	FFASIIY	FLGTSRHS	ISVGI	FGVLC	LMIG	ETVDREL	QKAGYDN	AHSA
S26A1_HUMAN	YSLYTS	FFANLIY	FLMGTSRHS	ISVGI	FSLLC	LMVGS	QVVDREL	QLAGPDP	SQDG...LQPGANS

S26A4_HUMAN	170	180	190	200	210	220	230		
S26A4_HUMAN	GTVL	NTTMTIDA	ARDTARV	LIASALT	LLVGI	ILIFGLQ	IGFIVRY	LADP	YVGGFTTAAAFQVL
S26A4_RAT	GSTL	NTTTLDT	GTRDAAR	VLLAST	LLVGI	ILVFGGLQ	IGFIVRY	LADP	YVGGFTTAAAFQVL
S26A4_MOUSE	GSAL	NSTTLDT	GTRDAAR	VLLAST	LLVGI	ILVFGGLQ	IGFIVRY	LADP	YVGGFTTAAAFQVL
S26A5_HUMAN	ATNG	EARDALR	VKVAMSV	TLLSG	IIICFLG	VCRFGF	VAVI	YLTBP	YVGGFTTAAAVHF
S26A3_HUMAN	LGLP	NNNSLLD	DERVRA	AAAAV	TVLSG	IIILAFGL	LRIGF	VVIY	YLSSESLSGGFTTAAAVHF
S26A6_HUMAN	TARDA	ARVQV	ASTLS	VLVGL	FLPQ	VGLGLI	HF	GFVVT	YLSSEPLVRGGFTTAAAVQVF
S26A2_HUMAN	DRI.....	CDKSCYA	IMVGS	TVTFI	AGVY	QVAMG	FFQV	GFVSV	YLSDALSGFVTGASPTIL
S26A1_HUMAN	AMLD.....	CGRDCYA	IRVATA	LTLMT	GLYQ	VLMLG	VLRL	GFVSA	YLSQPLDGFAMGASVTIL

S26A4_HUMAN	240	250	260	270	280	290	300					
S26A4_HUMAN	VSQKIL	VLNVST	TKNYNG	VLISII	YTLVEI	FPONIG	DTNLAD	FTAGLLT	IVVCM	AVKELN	DRFRHK	IPVPI
S26A4_RAT	VSQKIL	VLNVST	TKNYNG	VLISII	YTLVEI	FPONIG	DTNLAD	FTAGLLT	IVVCM	AVKELN	DRFRHK	IPVPI
S26A4_MOUSE	VSQKIL	VLNVST	TKNYNG	VLISII	YTLVEI	FPONIG	DTNLAD	FTAGLLT	IVVCM	AVKELN	DRFRHK	IPVPI
S26A5_HUMAN	TSMLK	YLFV	GKTKR	YSIF	SVVYST	VAVIQ	NVKNL	NVCSL	GVLMV	PGLLLG	GKFN	ERFKEKLP
S26A3_HUMAN	VSQKIL	FIFQ	LTVPS	HTDP	VSIF	KVLYS	VSFQ	IKTNIA	DLVTA	LIVLV	VVSI	YKIN
S26A6_HUMAN	VSQKIL	YVFG	LHLSS	HSGP	LSLI	YTVLEV	CWKLP	QSKVG	TVVTA	AVAGV	VLVVK	LLNDK
S26A2_HUMAN	TSQAK	YLLGL	NLPRT	NGVSL	ITTWI	HVFRN	IHKT	NL	CDLIT	SLLCL	LVLLP	TELNEH
S26A1_HUMAN	TSQKIL	LLGV	RIPRH	QGPG	MVVLT	WLSL	LRGAG	QANV	CDVVT	STVCL	AVLLA	KELSD

S26A4_HUMAN	310	320	330	340	350	360	370						
S26A4_HUMAN	EVIVTII	IATAIS	SYGANL	EKNYN	AGIVK	SIPR	GFLLP	PEL	FPVSLF	SEMLA	ASFIA	VVAIA	IAVSVGKVYA
S26A4_RAT	EVIVTII	IATAIS	SYGANL	EKNYN	AGIVK	SIPR	GFLLP	PEL	FPVSLF	SEMLA	ASFIA	VVAIA	IAVSVGKVYA
S26A4_MOUSE	EVIVTII	IATAIS	SYGANL	EKNYN	AGIVK	SIPR	GFLLP	PEL	FPVSLF	SEMLA	ASFIA	VVAIA	IAVSVGKVYA
S26A5_HUMAN	EFFAVV	MGTGIS	AGPNL	KESYN	VDVVG	TLEPL	GLD	PPAN	EDTSL	FHLV	VDAIA	IAI	IVGFSVTISMAKTLA
S26A3_HUMAN	EFIMT	VIAAGV	SYGCD	FKNRF	KVAVV	GDMM	NPGF	PPITE	DDVET	FQNT	VGD	CPG	IAMVAF
S26A6_HUMAN	ELLTL	LIGAT	GISY	GMLK	KHREF	VDVVG	NIPAG	LVPV	PAENT	QLFS	KLVGS	APTIA	VVGFAIAISLGIKIFA
S26A2_HUMAN	ELVVV	VVATL	ASHF	GKLH	ENYNS	SIAGH	PTG	FMP	PKV	EWNL	IPSV	AVDAIA	ISIGFAITVLS
S26A1_HUMAN	ELLVIV	VATLV	SHF	GQLH	KRFR	GSSV	AGDI	PTG	FMP	PKV	EWNL	IPSV	AVDAIA

S26A4_HUMAN	380	390	400	410	420	430	440						
S26A4_HUMAN	TKYDY	TIDGN	QEFIA	FGISN	IFSG	FFSCF	VATTAL	SRTAV	QESTG	GKTO	VAGI	ISAAI	VMIAIL
S26A4_RAT	TKHDY	IIDGN	QEFIA	FGISN	IFSG	FFSCF	VATTAL	SRTAV	QESTG	GKTO	VAGI	ISAAI	VMIAIL
S26A4_MOUSE	TKHDY	IIDGN	QEFIA	FGISN	IFSG	FFSCF	VATTAL	SRTAV	QESTG	GKTO	VAGI	ISAAI	VMIAIL
S26A5_HUMAN	NKHGY	QVDGN	QELIA	GLCNS	IGSL	FQTF	SISCS	LSRSL	VQESTG	GKTO	LAGC	LASL	MLLVIA
S26A3_HUMAN	LKYDY	PLDGN	QELIA	GLCNS	IGSL	FQTF	SISCS	LSRSL	VQESTG	GKTO	LAGC	LASL	MLLVIA
S26A6_HUMAN	LRHGY	RVDSN	QELIA	GLCNS	IGSL	FQTF	SISCS	LSRSL	VQESTG	GKTO	LAGC	LASL	MLLVIA
S26A2_HUMAN	KKHGY	TVKAN	QEMIA	GFPCN	IIPSF	FFHC	TTSA	ALAKT	LKESTG	GCHT	LSGV	VVAT	LVLLV
S26A1_HUMAN	RSHGY	SVRAN	QELIA	GLCNS	IGSL	FQTF	SISCS	LSRSL	VQESTG	GKTO	LAGC	LASL	MLLVIA

3. Risultati

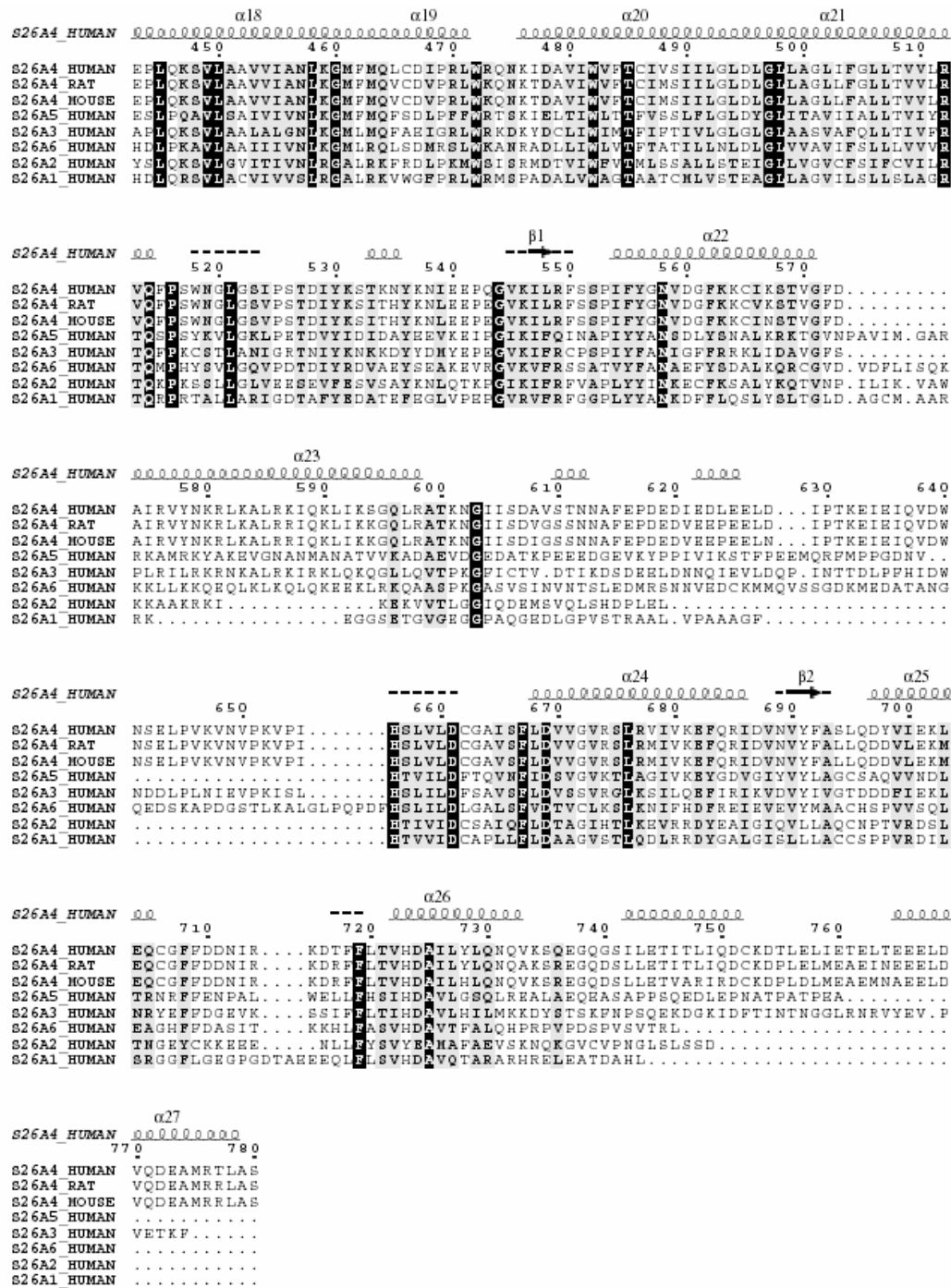


Fig. 3.1 Allineamento multiplo di Pendrina ed alcune sequenze omologhe rappresentative. La struttura secondaria predetta è indicata sopra alla sequenza con alfa eliche come spirali e beta strand come frecce.

posizione basale rispetto agli eucarioti. Tutte le sequenze del dominio STAS dei vertebrati, ad eccezione di quelle dei membri SLC26A11 (A11), assumono una posizione distale rispetto ai nematodi, funghi e piante. Le sequenze del dominio STAS delle piante hanno come sister group le sequenze del dominio STAS dei

membri A11, questo dato è a conferma di dati riportati precedentemente in letteratura dove è stato osservato un elevato livello di identità di sequenza del membro A11 umano con le sequenze dei trasportatori delle piante. Una particolare posizione nella topologia ottenuta è quella delle sequenze del dominio STAS di insetti. Queste infatti si presentano come sister group delle sequenze del dominio STAS dei trasportatori della famiglia SLC26A8 (A8) e come gruppo basale al raggruppamento delle sequenze del dominio STAS dei trasportatori delle famiglie SLC26A1 (A1) e SLC26A2 (A2).

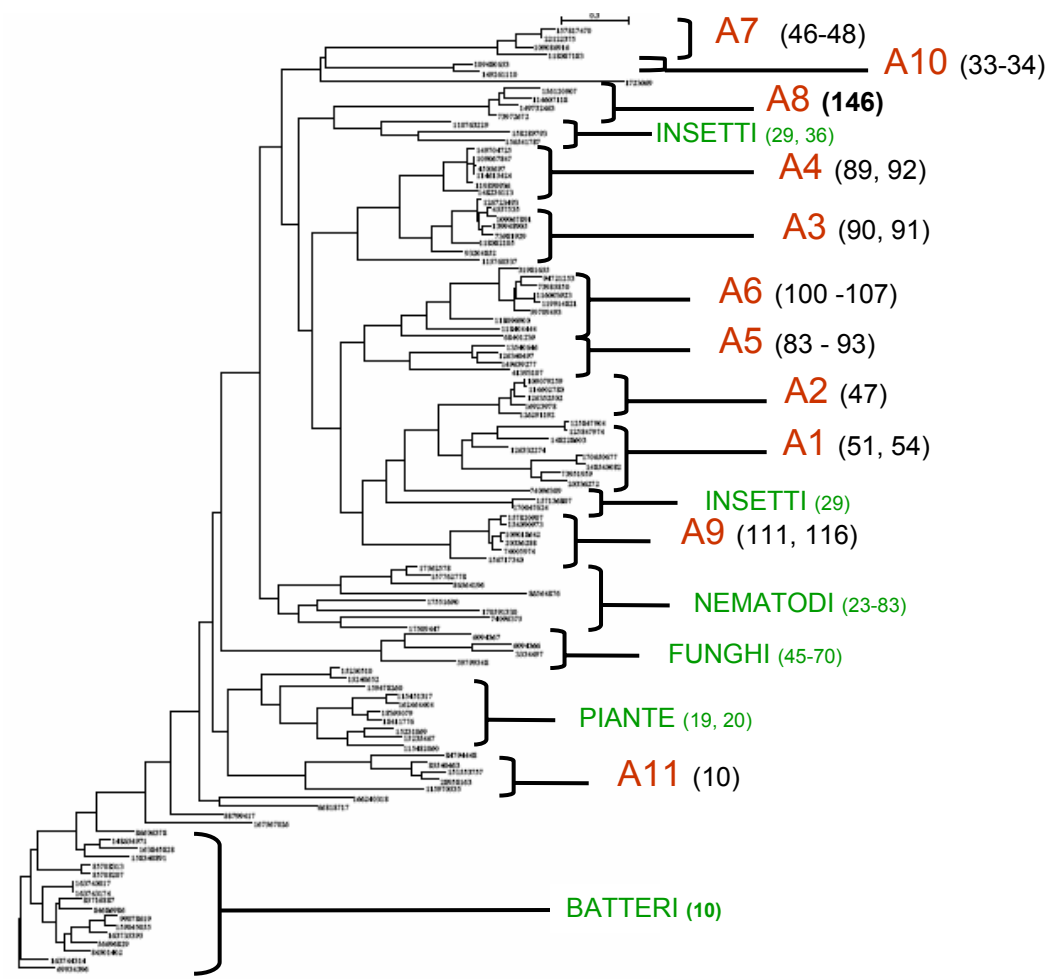


Fig. 3.2 L'albero Maximun Likelihood (ML) di sequenze del dominio STAS dei trasportatori del solfato SLC26A. I numeri tra parentesi indicano la lunghezza del loop variabile.

In ciascuna sequenza del multiallineamento è stata calcolata la lunghezza del loop variabile del dominio STAS e il valore calcolato è stato riportato sull'albero. Questo ha permesso di osservare la distribuzione delle dimensioni del loop lungo l'albero filogenetico. Si nota come gli organismi in posizione basale presentano un loop di

dimensioni inferiori agli altri gruppi di sequenze (batteri: 6aa; piante: 15aa). Nelle sequenze del dominio STAS di funghi e di nematodi il loop assume lunghezze che variano da 15 a 75 residui mentre nei vertebrati assumono lunghezze maggiori. Il gruppo di sequenze del dominio STAS di A11 presenta una dimensione del loop di 6 residui, simile a quella del gruppo basale dei batteri, in accordo con quanto precedentemente descritto. Le sequenze del dominio STAS con posizione più distale (A9; A5; A6) sono tra quelle che presentano il loop più lungo. Isolato il caso delle sequenze di A8 il cui loop arriva a 142 residui. Più complesso risulta quello che si osserva nella distribuzione della dimensione del loop nelle sequenze del dominio STAS degli insetti, in cui il loop si mantiene costituito da 29 residui nei due gruppi di sequenze. Un più ampio campionamento di dati potrà in futuro chiarire tale distribuzione. Si può comunque già notare come sia la lunghezza del loop variabile a distinguere principalmente le sequenze del dominio STAS dei trasportatori SLC26A.

3.2 Estremità N-terminale

Esperimenti di immunolocalizzazione hanno dimostrato che l'estremità N-terminale è localizzata nel citosol (Royaux et al., 2000; Gillam et al., 2004). La regione corrispondente include 82 residui che precedono la prima elica transmembrana in cui l'analisi, condotta con i metodi per la predizione 1D, ha permesso di individuare una sequenza iniziale disordinata di 20 residui a cui seguono tre α -eliche. L'applicazione di PSORT esclude la presenza di sequenze segnale per la localizzazione intracellulare della proteina mentre SignalP rileva con bassa probabilità la presenza di un sito di taglio tra la posizione 28 e 29. La ricerca di motivi conservati con ELM ha permesso di individuare due potenziali siti di legame per domini SH3, in posizione 8-14 e 19-25, e un sito TRG_endocytic_2 a livello dei residui 27-30. Le proteine che contengono un dominio SH3 legano peptidi PXXP *proline-rich* e sono coinvolte in molti processi biologici tra cui: traduzione del segnale, organizzazione del citoscheletro, traffico delle proteine di membrana, assemblamento di organelli. I siti TRG_endocytic_2 sono segnali di uscita basati sulla tirosina responsabili delle interazioni con le subunità *mu* di complessi di proteine adattatrici e sono coinvolti nel traffico intracellulare delle proteine di membrana. Questi segnali determinano la via con cui una proteina viene trasportata e quindi la loro destinazione finale. Il motivo YXXphi è stato trovato nelle estremità citosoliche di molte proteine di membrana come segnale di endocitosi e dirige il traffico verso gli endosomi e alla via di secrezione. Altri due siti

identificati nelle sequenze 47-50 e 72-75 rappresentano motivi di legame a domini PDZ (LIG_PDZ_3). I domini PDZ sono presenti in proteine coinvolte nella trasduzione del segnale e hanno un ruolo fondamentale nelle interazioni per l'assemblamento di complessi multiproteici ma i siti di legame si trovano più spesso nelle estremità C-terminali delle proteine *target*. All'estremità N-terminale inoltre vengono predetti anche potenziali siti di fosforilazione per diversi tipi di kinasi. A livello dei residui 16-23 e 42-49 ELM individua due siti per kinasi GSK3, che generalmente fosforilano le proteine *target* inibendole. In corrispondenza di una di queste sequenze vengono inoltre indicate come potenziali siti di fosforilazione le sequenze ERKTLRE (residui 42-48) e LRESLAK (residui 46-52). In questo caso la Kinasi coinvolta è cAMP dipendente (PKA). L'analisi della sequenza relativa all'estremità N-terminale con i programmi NetNGlyc e NetOGlyc ha permesso di escludere la presenza di potenziali siti di glicosilazione in questa regione.

3.3 Regione transmembrana

La regione transmembrana consta di circa 429 aminoacidi che includono il dominio dei trasportatori del solfato (PF00916) e due motivi conservati, il motivo per la famiglia SLC26A (PS01130) e il motivo di Saier a livello dei residui 382-384.

3.3.1 Predizioni 1D

Dalla predizione di struttura secondaria con Porter e Psipred risulta che la regione transmembrana è prevalentemente costituita da α -eliche, gran parte delle quali si presume corrispondano ai segmenti di membrana. In posizione 242-243 e 250-255 Psipred predice β -strand. Dal server Conseq sono stati ricavati i risultati di predizione dell'accessibilità al solvente per ogni residuo della sequenza di Pendrina e il disordine è stato riscontrato nelle regioni comprese tra i residui 130-140, 155-175, 285-295, 329-334 e 373-380.

Predizione della topologia di membrana

La sequenza di Pendrina è stata analizzata con i metodi di predizione per le regioni transmembrana più avanzati, alcuni dei quali permettono di imporre la localizzazione di parti della proteina o di residui specifici in un determinato compartimento. Nel caso di Phobius e HMMTOP v2.1 le estremità N e C-terminale sono state limitate al citosol. La maggior parte dei metodi utilizzati predice 12 segmenti transmembrana ma in alcuni casi vengono predette

3. Risultati

regioni di membrana laddove altri indicano la presenza di un loop (Tab. 3.1).

A livello di tre regioni in particolare, solo alcuni metodi predicono potenziali segmenti transmembrana: regioni tra 161 e 180, 235 e 270, 325 e 340 (Fig. 3.3). Queste regioni sono state quindi valutate per le caratteristiche nella composizione aminoacidica e con i dati relativi alla predizione di struttura secondaria, del disordine e accessibilità al solvente utilizzando per il confronto l'editor Jalview. Nella regione compresa tra 161 e 180, Disopred predice disordine e Psipred struttura secondaria di tipo coil. In corrispondenza della regione 235-270, Psipred predice un β -strand e alcuni aminoacidi risultano non accessibili al solvente. Nella regione compresa tra 325 e 340 i metodi di predizione di struttura secondaria rilevano un lungo tratto con struttura a coil. In questa regione sono presenti anche diverse proline ravvicinate.

Programma	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q
MEMSAT3	83- 110	113- 136	140- 157		184- 211	214- 238		264- 287	300- 319		349- 376		409- 436	439- 466	490- 511
HMMTOPv.2	83- 100	107- 131	138- 156	161- 179	186- 204	219- 236		267- 285	300- 319		350- 368	385- 407	422- 439	448- 466	479- 503
HMMTOPv.1	84- 114	110- 130	136- 155	161- 180	186- 206	212- 232		264- 286	299- 319	325- 345	351- 371	386- 407	422- 442	448- 468	483- 507
HMMTOP constrained	83- 100	105- 129	138- 155		186- 210	219- 236		267- 285	296- 319		350- 368	385- 407	422- 439	448- 465	480- 504
PHOBIUS constrained		105- 130	138- 155		186- 212	218- 240	247- 264	270- 288	300- 317		345- 364	385- 407	422- 442	449- 465	480- 511
PoliPhobius	85- 100	102- 130	136- 156		187- 215	221- 240		267- 286	297- 317		346- 371	385407	421- 442	448- 466	480- 511
TMHMM v.2	83- 104	109- 131			186- 208	218- 240			296- 318		350- 372	385- 407	422- 439		480- 502
TMHMM V.1	83- 104	109- 131	138- 156		187- 208	219- 240		267- 285	296- 316		350- 372	385- 407	422- 439		480- 502
PRODIV- TMHMM	88- 108	113- 133	136- 156		184- 204	220- 240		264- 284	300- 320		350- 370	386- 406	422- 442	445- 465	490- 511
TOPMOD	88- 109	111- 132	136- 156		185- 206	215- 236		267- 288	295- 317		345- 366	385- 407	421- 442	444- 465	490- 511
CONSENSO	83- 108	113- 130	136- 156		186- 212	218- 236		267- 285	300- 319		350- 368	385- 407	422- 442	448- 465	480- 511
TM	1	2	3		4	5		6	7		8	9	10	11	12

Tab. 3.1 Regioni da A a Q individuate come possibili segmenti transmembrana dall'analisi di predizione della topologia di membrana di Pendrina con i diversi metodi utilizzati e consenso ottenuto riportando i limiti maggiormente rappresentati nelle diverse predizioni.

Regione 161 - 180

	10	20	30
<i>sequence/153-188</i>	LSMAPDEHFLVSSSNGTVLNTTMI	DTAARDTARVLI	
<i>accessibilita/153-188</i>	eeeeeeeeeeeeeeeeeeeeeeeeeeee	eeeeeeeeeeeeeeeeeeeeeeeeeeee	ebbb
<i>Disopred/40-56</i>	- - - DDDDDDDDD - - DDDDDDD - - D - - - - - - - -		
<i>top-mod/153-188</i>	MMMMLLLLLLLLLLLLLLLLLLLLLLLL	LLLLLLLLLLLLLLLLLLLLLLLLLLLL	MMMM
<i>HMMTOP/153-188</i>	HHHHssssHHHHHHHHHHHHHHHHHH	HHHHHHHHHHHHHHHHHHHHHHHHHH	nnnnnnHHH
<i>spritz/153-188</i>	HHCCCCCHHHHHHHHHHHHHHHHHHH	HHCCCCCHHHHHHHHHHHHHHHHHHH	
<i>PSIPRED/153-188</i>	HHHCCCCCCCCCCCCCCCCCCCCCCCCC	CCCCCCCCCCCCCCCCCCCCCCCCCH	HHHHHHH

Regione 235 - 270

	10	20	30
<i>sequence/235-270</i>	QLKIVLNVSTKNYNGVLSI	IYTLVEIFQNI	GDTNLA
<i>accessibilita/235-270</i>	ebbbbeeeeeeeeeeeeeebbbb	beebbeebbeebbeebbeeb	beebbeeb
<i>Disopred/57-56</i>	- -		
<i>phobius/235-270</i>	nnnnnnnnnnnnnnHHHHHHHHHHHHHHHHHH	ssssss	
<i>top-mod/235-270</i>	MMLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	LLLLLLLLLLLLLLLLLLLLLLLLLLLL	MMMM
<i>spritz/235-270</i>	HHHHHCCCCCCCCCHHHHHHHHHHHHHHHHH	CCCCCHH	
<i>PSIPRED/235-270</i>	HHHHHCCECCCCCEEEEEHHHHHHHHH	CCCCCHH	

Regione 325 - 340

	10	20	30
<i>sequence/313-349</i>	ISYGANLEKNYNAGIVKS	IPRGFLPPELP	PPVSLFSEM
<i>accessibilita/313-349</i>	bbbbbeeeeeebbeebbeeb	beebbeebbeebbeebbeeb	beeb
<i>Disopred/63-63</i>	- - - - - D -		
<i>top-mod/313-349</i>	MMMMLLLLLLLLLLLLLLLLLLLLLLLLLLLL	LLLLLLLLLLLLLLLLLLLLLLLLLLLL	MMMM
<i>HMMTOP_v.1/313-349</i>	nnnnnnnnnnnnnnHHHHHHHHHHHHHHHHHH	ssssssssss	
<i>spritz/313-349</i>	HHHHHHHCCCCCCCCCECCCCCCCCCCCCCCC	CCCCCHHHHHH	
<i>PSIPRED/313-349</i>	HHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCCCCHHHHHH	

Fig. 3.3 Confronto tra le predizioni di accessibilità, disordine e struttura secondaria per la valutazione delle regioni predette come transmembrana da pochi metodi. e: accessibile al solvente; b: buried; D: presenza di disordine. TOPMOD: M: predizione di residuo transmembrana; L: loop. HMMTOP e Phobius: n: loop intracellulare; s: loop extracellulare. Struttura secondaria, Spritz e PSIPRED: H: elica; C: coil; E: β -strand.

Utilizzando il consensus di almeno 5 metodi sono stati stabiliti il numero e i limiti dei segmenti transmembrana (Tab 3.1). Questo approccio permette di ottenere una migliore sensibilità di predizione anche se le regioni di dubbia predizione vengono automaticamente escluse. Nelle successive fasi di analisi queste regioni sono state però studiate con altri metodi per verificare la presenza di regioni rientranti o eliche interfaccia.

Composizione aminoacidica

L'analisi della sequenza di Pendrina dal punto di vista aminoacidico permette di identificare due regioni ricche di residui di tirosina e Triptofano in corrispondenza della seconda elica transmembrana e nel loop che collega l'ottava e la nona elica transmembrana. Entrambe le regioni presentano molti residui di piccole dimensioni come Glicina e Alanina. Residui di Prolina sono presenti nella 2° e nella 3° elica transmembrana. La prolina in posizione 123 è quella maggiormente conservata mentre le proline in posizione 140 e 142 risultano conservate in almeno il 10% delle sequenze, nelle altre la prolina è stata sostituita con residui di piccole dimensioni tra cui la più frequente è l'alanina. Residui carichi conservati o cluster di residui carichi sono stati individuati a livello dei loop ma anche all'interno di alcuni segmenti transmembrana (vedi Fig. 3.4).

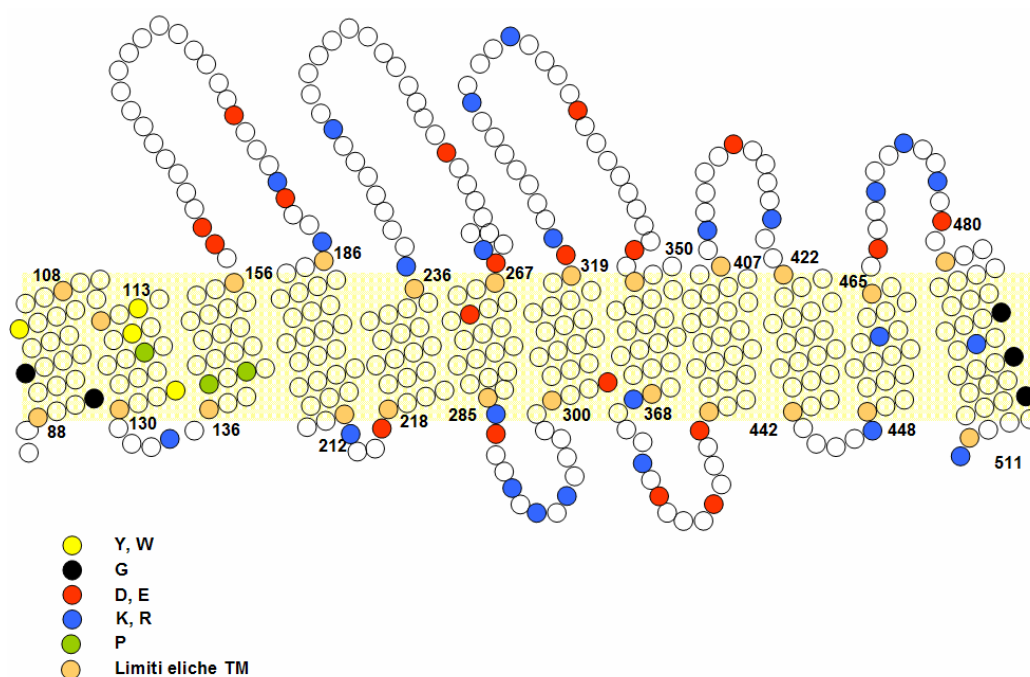


Fig. 3.4 Schema topologico delle regioni transmembrana di Pendrina. Gli aminoacidi sono colorati secondo le loro caratteristiche fisico-chimiche e i limiti di ogni regione indicati. L'N-terminale è a sinistra, il C-terminale a destra nella figura.

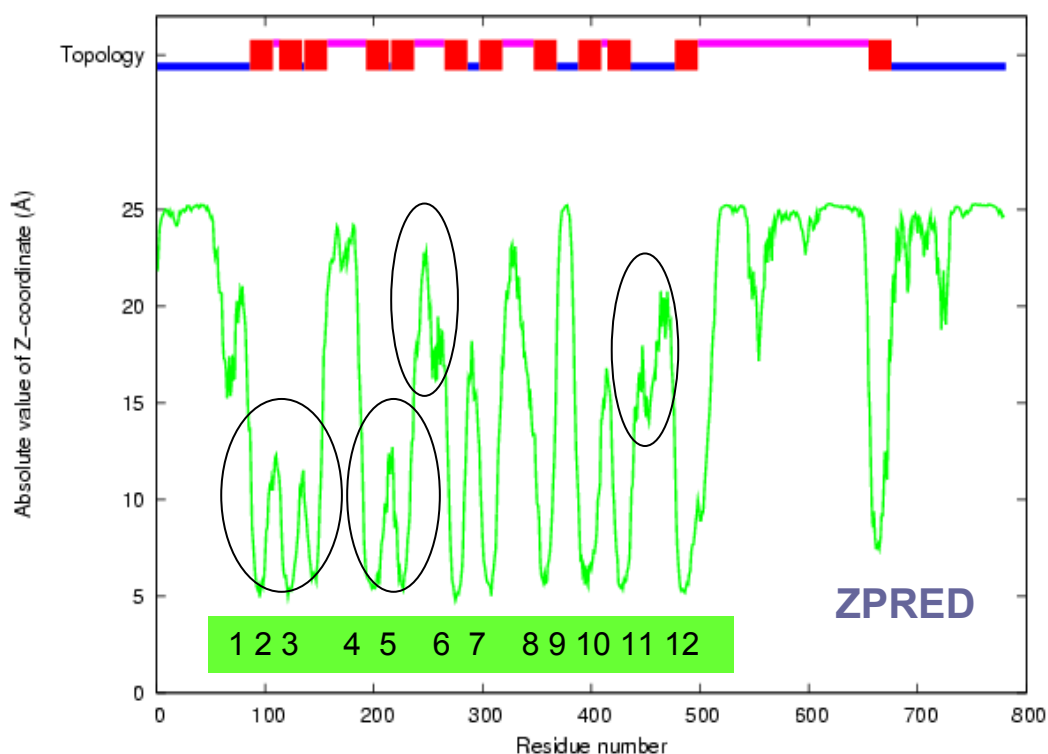
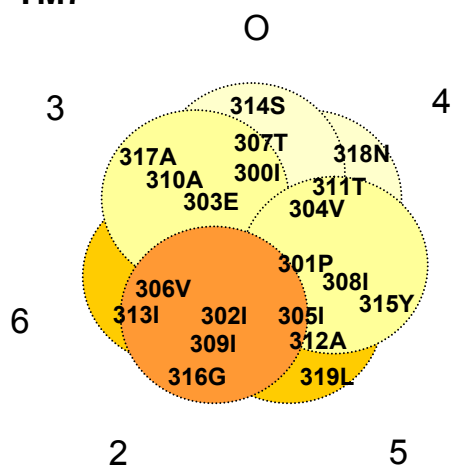


Fig. 3.5. Output dell'analisi della sequenza di Pendrina ottenuto da ZPRED. Il grafico mostra la distanza in Angstrom di ciascun residuo dal centro della membrana (0 Å). I valori ottenuti dal programma hanno generalmente un errore di 3.6Å, infatti nessuna elica predetta raggiunge il valore 0. ZPRED si basa sulla predizione di PRODIV-TMHMM per individuare le regioni transmembrana di una sequenza proteica. In questo caso PRODIV-TMHMM non prevede l'elica 11 e individua un'altra regione transmembrana in corrispondenza dell'estremità C-terminale. I numeri in basso su sfondo verde indicano i segmenti transmembrana. Sono stati segnati con un cerchio le tre putative regioni rientranti.

Ricerca delle regioni rientranti o eliche interfaccia

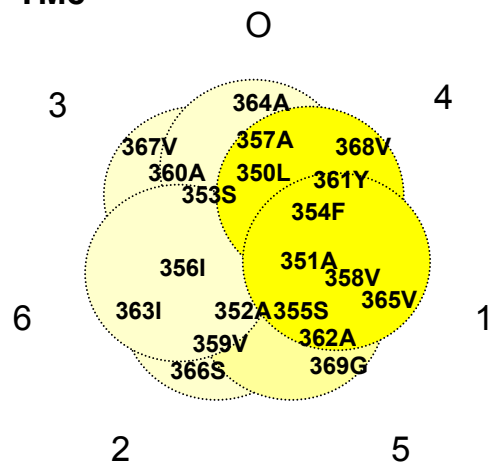
Per la ricerca di regioni rientranti o eliche interfaccia sono stati utilizzati i programmi TOPMOD e ZPRED. TOPMOD non ha riscontrato la presenza di regioni rientranti mentre dai dati ottenuti da ZPRED è possibile individuare alcune regioni di membrana che non attraversano completamente il doppio strato lipidico (Fig. 3.5). Infatti, l'elica 2 e le eliche 4 e 5 si allontanano dal centro della membrana di circa 5Å e poi proseguono direttamente con le eliche successive. I residui dell'elica 11 inoltre non raggiungono la distanza di 0Å che corrisponde al centro della membrana ma rimangono a una distanza di circa 15Å che corrisponde all'interfaccia tra la membrana e l'ambiente acquoso.

TM7



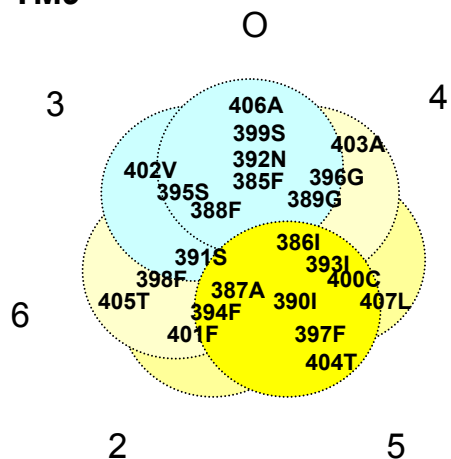
300-IPIEVIVTIIATAISYGANL-319

TM8



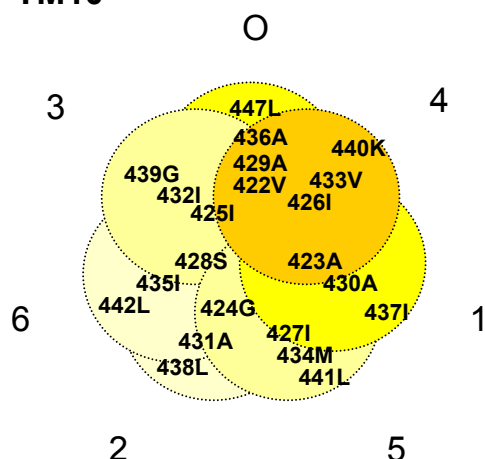
350-LAASFSAIVVAYAIASVVG-368

TM9



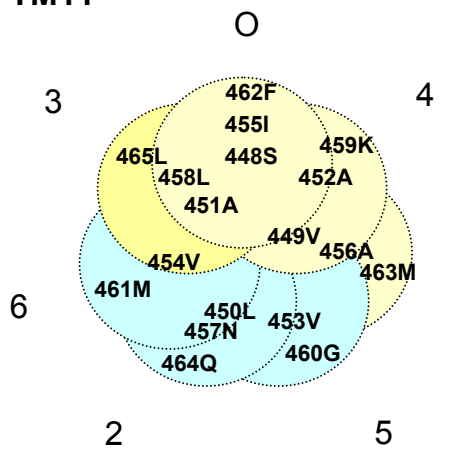
385-FIAFGISNIFSGFFSCFVATTAL-407

TM10



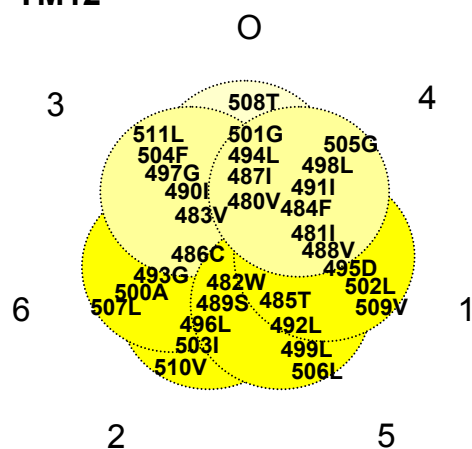
422-VAGIISAAIVMAILALGKLL-442

TM11



448-SVLAADVIANLKGFMFQL-465

TM12



480-VIWFVFTCVISILGLDLGLLAGLIFGLLTVVL-511



Fig. 3.6 Grafico in cui vengono riportati i punteggi assegnati alle sette facce di ogni elica transmembrana dal server LIPS per classificare le eliche di una proteina politopica in base alla loro accessibilità ai lipidi.

Predizione dell'orientamento delle eliche nella membrana

Il programma LIPS ha permesso di calcolare per ogni faccia di ciascuna elica la propensione ad orientarsi verso l'ambiente idrofobico dei lipidi di membrana. I risultati sono stati rappresentati con uno schema che riporta per ogni elica la disposizione dei residui sul modello a sette facce ricavato dalle strutture coiled-coil e le facce, indicate con un cerchio, sono state colorate con intensità e tonalità diverse in base al valore ottenuto da LIPS (Fig. 3.6).

Risulta da questi dati che le eliche 4, 6, 7, 10 sono quelle che presentano almeno una faccia con forte propensione ai lipidi mentre le eliche 1, 2, 3, 5, 9 e 11 presentano facce con minor accessibilità ai lipidi che probabilmente sono implicate nelle interazioni tra eliche. Con il server RANTS è stato possibile classificare le eliche in base alla loro accessibilità al solvente e risulta che le eliche 6 e 7 presentano un punteggio maggiore che indica un'elevata accessibilità ai lipidi mentre le eliche 1 e 5 potrebbero essere inglobate all'interno dei riarrangiamenti tra eliche (Fig. 3.7).

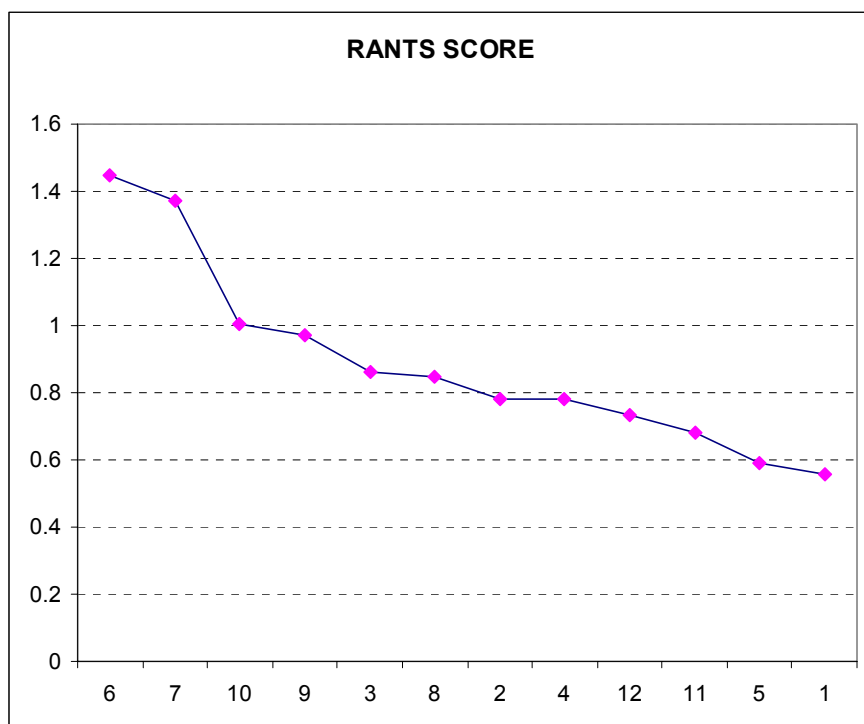


Fig. 3.7 Grafico in cui vengono riportati i punteggi assegnati a ciascuna elica dal server RANTS che utilizza la scoring function di LIPS per classificare le eliche di una proteina politopica in base alla loro accessibilità ai lipidi.

Ricerca di motivi GXXXG

Nelle eliche 1, 2, 4, 5, 7, 8 e 12 è possibile individuare motivi GXXXG o motivi simili in cui la glicina è sostituita da un altro residuo di piccole dimensioni come l'alanina. In particolare nella 12° elica si distinguono tre motivi GXXXG successivi. L'elica 1 presenta due motivi, il motivo tipico GXXXG in posizione 91-95 e il motivo degenerato AXXXG in posizione 98-102. Le glicine 91, 95, 102 e l'alanina 98 sono localizzate sulla faccia n° 3 meno esposta ai lipidi. L'elica 2 presenta un motivo degenerato GXXXA in posizione 116-120 in cui la glicina e l'alanina sono localizzate sulla faccia n°3 meno esposta ai lipidi. L'elica 4 presenta un motivo GXXXG in posizione 205-209 le cui glicine si trovano invece sulla faccia n°5 maggiormente esposta ai lipidi. L'elica 5 presenta un motivo degenerato GXXXA in posizione 222-226 in cui la glicina e l'alanina sono localizzate rispettivamente sulle facce n°4 e n°1 che presentano una diversa propensione per l'ambiente idrofobico. L'elica 7 presenta un motivo degenerato AXXXG in posizione 312-316 in cui la glicina e l'alanina sono disposte sulla faccia n°2 con forte propensione per i lipidi. L'elica 8 presenta un motivo degenerato AXXXG in posizione 364-368 in cui l'alanina e la glicina sono disposte sulla faccia n° 4 maggiormente esposta ai lipidi. A livello dell'elica 12 si trovano tre motivi GXXXG consecutivi in cui le glicine in posizione 497, 501 e 505 sono localizzate sulla faccia n°0 meno esposta ai lipidi mentre la glicina 493 si trova sulla faccia 6 maggiormente esposta all'ambiente idrofobico (Tab. 3.2).

MOTIVO GXXXG	
TM 1	91-GXXXG-95, 98-GXXXA-102
TM 2	116-GXXXA-120
TM 4	205-GXXXG-209
TM 5	222-GXXXA-226
TM 7	312-AXXXG-316
TM 8	364-AXXXG-368
TM 12	493-GXXXGXXXGXXXG-505

Tab.3.2 Motivi GXXXG o G/AXXXG/A riscontrati nelle eliche transmembrana predette. G: Glicina; X: qualsiasi aminoacido.

3.3.2 Predizione 3D

Ricerca del template

La consultazione delle banche dati di strutture proteiche tramite Interpro ha permesso di individuare similarità della sequenza di Pendrina con domini noti. L'output di Interpro fornisce uno schema topologico in cui i domini identificati per le diverse banche dati di proteine note vengono posizionati sulla sequenza della proteina target. Per ogni dominio è possibile accedere attraverso un *link* alle informazioni disponibili della banca in cui è stato annotato il dominio. Le banche forniscono anche informazioni relative alle strutture proteiche che mostrano una certa similarità con la proteina target e spesso contengono *link* che permettono di consultare le informazioni relative a queste strutture delle banche di strutture proteiche come SCOP e CATH. Superfamily Search per esempio restituisce una pagina con la classificazione strutturale e funzionale della famiglia secondo SCOP. Dalla banca dati ModBase è stato possibile ottenere informazioni riguardo a un possibile template da utilizzare per la predizione di struttura terziaria di Pendrina. Esiste infatti la struttura cristallografica di una proteina transmembrana di Citocromo C ossidasi di *Paracoccus denitrificans* (PDB code 1ar1) che mostra un'identità di sequenza del 14% con Pendrina lungo la regione 101-516, corrispondente a gran parte della regione transmembrana. La Citocromo C-ossidasi è classificata in CATH nella classe di proteine Mainly alpha, con architettura *updown bundle* e nella banca dati di proteine transmembrana è raggruppata tra le proteine con 12 segmenti transmembrana. La Citocromo C-ossidasi contiene Rame e svolge la sua funzione nei mitocondri dove permette la riduzione dell'ossigeno ad acqua, accompagnando la reazione con l'estruzione di 4 protoni nel compartimento mitocondriale.

Date le differenze funzionali di questa proteina con Pendrina si è voluto verificare la loro possibile similarità strutturale con un programma di *fold recognition*. La ricerca con Metaserver ha dato come risultato la stessa struttura del Citocromo C-ossidasi anche se il punteggio di 3D-Jury è di 34.9, un valore inferiore a quello indicato per considerare affidabile il risultato della ricerca.

Allineamento target- template

Si è voluto comunque provare a costruire un allineamento tra le due proteine allo scopo di verificare che ci sia una certa corrispondenza tra il riarrangiamento delle eliche in Pendrina e quello in Citocromo C-ossidasi. Per la costruzione dell'allineamento

sono stati utilizzati ALIGN e SSEA. L'allineamento locale, utilizzando la modalità profilo contro profilo, ottenuto con ALIGN esclude le prime tre eliche di Pendrina mentre un allineamento globale interrompe alcune eliche all'interno della sequenza di Pendrina. Il programma SSEA permette di mantenere insieme le eliche predette ma il confronto risulta difficoltoso perché nella Citocromo C-ossidasi sono presenti tratti transmembrana scomposti che non attraversano con un unico tratto la membrana e che quindi vengono allineati con eliche diverse in Pendrina. Una valutazione manuale dell'allineamento ottenuto ha infine permesso di escludere la possibilità di costruire un modello tridimensionale di Pendrina utilizzando come templatato la struttura del Citocromo C-ossidasi, in quanto troppo poco affidabile.

Ricerca di motivi lineari

La ricerca di pattern funzionali di sequenza, di interazioni con ligandi o di eventuali modifiche post-traduzionali vengono generalmente svolte mediante i database PROSITE o ELM. I motivi identificati da ELM all'interno della regione transmembrana sono stati considerati automaticamente come falsi positivi data l'inaccessibilità delle regioni che attraversano la membrana.

NetNGlyc individua tre potenziali sequenze Asn-Xaa-Ser/Thr, in cui l'asparagina potrebbe essere glicosilata, a livello delle posizioni 167, 172 e 241. I residui 167 e 172 si trovano nel *loop* tra l'elica 3 e 4 mentre il sito 241 è localizzato nel *loop* tra l'elica 8 e 9. Entrambe le regioni dovrebbero essere esposte all'ambiente extracellulare.

3.4 Estremità C-terminale

L'estremità C-terminale è costituita dai 269 aminoacidi che si estendono nel citosol. I primi 23 residui rappresentano la regione che collega l'ultimo segmento transmembrana con il dominio STAS. Quest'ultimo comprende la regione dal residuo 535 al residuo 729 e infine gli ultimi 51 aminoacidi costituiscono la coda terminale della proteina. La ricerca di domini all'estremità C-terminale ha dato risultati positivi per il dominio STAS annotato in tre banche diverse: Superfamily sequence search (SSF5209), Pfam (PF01740) e PROSITE (PS50801). Superfamily Sequence Search individua il dominio STAS, nella regione tra i residui 532 e 725 mentre Pfam (PF01740) lo localizza tra i residui 536 e 725 e PROSITE tra i residui 535 e 729. In base alla conservazione osservata in corrispondenza dei limiti di questa regione si è deciso di fissare la sequenza del dominio STAS in corrispondenza della regione compresa tra i residui indicati da

PROSITE. La struttura viene classificata nella classe di proteine alpha e beta (a/b), principalmente β -sheet paralleli, il fold è definito come SpollAA-like in cui viene riconosciuto un core proteico di 4 giri di una superelica alpha - beta.

3.4.1 Predizioni 1D

Struttura secondaria

Le predizioni di struttura secondaria di PSIPRED e Porter (Spritz) coincidono per la regione del dominio STAS mentre per la regione di linking e la coda terminale presentano delle discordanze. Nella regione di linking PSIPRED predice un β -strand iniziale seguito da una struttura a coil mentre Spritz rileva la presenza di brevi tratti di α -elica e β -strand alternati. Nella coda terminale Spritz predice tre α -eliche separate da brevi tratti a coil e la predizione di PSIPRED coincide solo con l'ultima di queste tre eliche.

Nella regione del dominio STAS vengono confermate le informazioni relative alla disposizione degli elementi di struttura secondaria noti per la struttura del dominio SpollAA (Fig. 3.8), ad eccezione del primo β -strand (1β) in Pendrina. Sulla sequenza di Pendrina si possono distinguere tre β -strand (2β , 3β , 4β , 5β) e quattro α -elica (1α , 2α , 3α , 4α) disposti secondo l'ordine $1\beta - 2\beta - 1\alpha - 3\beta - 2\alpha - 4\beta - 3\alpha - 4\alpha$. Tra la prima α -elica e il terzo β -strand si trova il loop variabile, che in Pendrina è lungo 84 aminoacidi, in cui viene predetta un' α -elica in posizione 575-596 e un β -strand in posizione 603-606. Il resto del loop ha una struttura a coil.

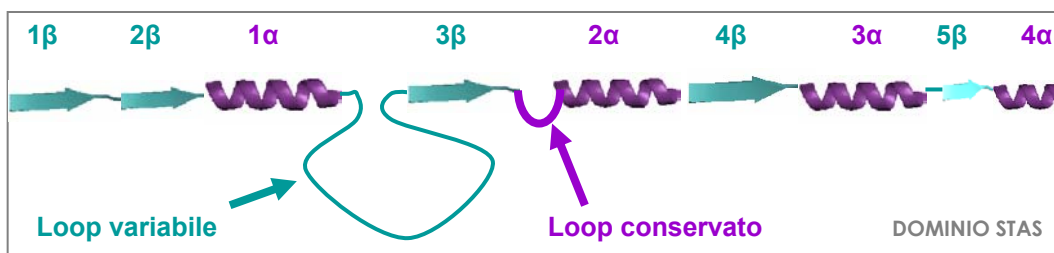


Fig. 3.8 Struttura secondaria del dominio STAS ricavata dal confronto fra le sequenze dei trasportatori del solfato e le struttura 3D di SpollAA. Si riconoscono quattro α -eliche (in viola) e quattro β -strand (freccie in azzurro), il quinto β -strand non è presente in tutte le strutture note provenienti da diversi organismi. Il loop conservato è presente anche nei trasportatori del solfato mentre il loop variabile presenta una sequenza che differisce tra i vari membri della famiglia SLC26A.

Predizione del disordine

Disopred e Spritz, anche se con diversa sensibilità, rilevano la presenza di disordine in particolare regioni: nella regione di linking subito dopo il β -strand; all'interno del loop variabile, sia nella parte centrale che in corrispondenza dell'elica predetta; tra l' α -elica 3a e il β -strand 5 β ; infine lungo tutta la regione della coda terminale di 50 aminoacidi.

Composizione aminoacidica

A livello dell'estremità C-terminale sono state individuate tre regioni con un'elevata incidenza di residui carichi. All'interno del loop variabile del dominio STAS si distinguono un cluster di cariche positive nella regione compresa tra i residui 575 e 601 (26aa) e, adiacente a questo, nella regione compresa tra i residui 616 e 643 (27aa), un cluster di cariche negative. Nella regione al C-terminale che si estende dal residuo 751 al 773 (22aa) si distingue un altro cluster di cariche negative.

PROGRAMMA	Probabilità di Coiled-coil in Pendrina		
	CCM-I (29-53)	CCM-II (570-590)	CCM-III (747-778)
COILS	+++	+	+++++
PCOILS	+	+	+++
NCOILS (PAT)	-	-	+++++
COILS 2 (NPS)	+	+++	+++++
MARCOIL	+	+	+++
MULTICOIL	+	+	+
Paircoil 2	-	-	+++++

Tab. 3.3 Predizione di CCM ottenute con diversi programmi. I valori di probabilità sono stati comparati secondo il seguente ordine : +++++ forte probabilità, +++ moderata, + debole, - nessuna.

Ricerca di motivi coiled-coil (CCM)

Dai risultati di PAT è stato possibile ottenere la predizione NCOILS che individua la presenza di una struttura coiled-coil nella regione compresa tra i residui 749 e 769. Per confermare questo risultato la sequenza di Pendrina è stata analizzata per la presenza di strutture coiled-coil utilizzando altri metodi disponibili in rete (Tab. 3.3). Quasi tutti i metodi, ad eccezione di Multicoil, predicono con buona probabilità una struttura coiled-coil nella coda dell'estremità C-terminale (CCM-III). Altre strutture coiled-coil vengono predette nell'estremità N-terminale nella regione compresa tra i residui 29 e

Ortologi di pendrina	Marcoil					
	CCM-I		CCM-II		CCM-III	
Organismi						
Homo	29-53	9.7%	574-594	3.4%	739-779	52.9%
Pongo abelii	29-52	9.4%	574-594	3.9%	732-776	52.9%
Pan	29-53	9.7%	574-594	3.4%	735-776	48.1%
Macaca	29-53	9.7%	574-594	3.9%	732-776	52.9%
Canis	29-52	8.8%	574-596	4.4%	736-776	21.8%
Rattus	/	/	574-595	3.8%	733-776	37.3%
Mus musculus	/	/	574-595	3.8%	756-772	14.2%
Monodelphis	31-43	12.2%	571-593	4.1%	731-762	2.6%
Xenopus leavis	/	/	565-576	1.3%	736-765	12.5%
Xenopus tropicalis	28-49	4.8%	574-589	2.9%	765-771	1.1%
Danio rerio	/	/	553-578	34.8%	/	/
Strongylocentrotus	33-59	5.4%	616-642	36.7%	/	/

Tab. 3.4 Posizione e punteggio dei motivi coiled-coil predetti mediante Marcoil nella sequenza di Pendrina e nei suoi ortologi. CCM-I localizzato nell'estremità N-terminale, CCM-II all'interno del loop variabile e CCM-III nella coda del C-terminale.

Nelle proteine della famiglia SLC26A i motivi coiled-coil CCM-II e CCM-III vengono predetti in corrispondenza delle stesse regioni di Pendrina. Il CCM-II è predetto con una bassa probabilità nei membri A2, A4 (Pendrina), A5, A8 di *Pan troglodita* e A9, mentre nei membri A3 e A6 viene predetto con buona probabilità. Il CCM-III è predetto con bassa probabilità nei membri A2, A5 di ratto e topo, A8 e A9 mentre con buona probabilità nei membri A1, A4, A5 di uomo e A11 di uomo. La presenza di una regione coiled-coil al C-terminale di Pendrina appare dunque come un adattamento di una parte dei trasportatori SLC26A. Questo potrebbe indicare un meccanismo molecolare diverso a seconda della famiglia specifica del trasportatore.

3.4.2 Predizione 3D

Per l'estremità C-terminale è stato possibile pensare di costruire un modello tridimensionale del dominio STAS. E' noto infatti che le strutture di proteine ASA, che mostrano lo stesso fold, sono state determinate sperimentalmente.

Ricerca del template

Dalla ricerca in banche dati di proteine con Interpro è stato possibile risalire alle informazioni relative al dominio STAS identificato in Pendrina. Le banche dati di PROSITE e PFAM permettono di ottenere una lista di possibili template da poter utilizzare per la costruzione del modello. Superfamily sequence search riporta la classificazione del dominio secondo SCOP e permette di accedervi direttamente con diversi link. Dalla ricerca in SCOP risulta che le strutture proposte come template sono di SpollAA di diversi organismi batterici (Tab. 3.5).

Template	PDB code	Tecnica (risoluzione)	Similarità	Metaserver score
SP2AA BACSH	1H4XA 1H4Y 1H4ZA	Xray (1.16) Xray (1.61) Xray (2.74)	20%	58.3
SP2AA BACST	1TH8B 1THNB 1TIDB 1TILB	Xray (2.4) Xray (2.5) Xray (2.5) Xray (2.7)	12.5%	62.5
SP2AA BACSU	1AUZA 1BUZA	NMR NMR	14.8%	65.5
Y1442 THEMA	1T6RA 1VC1A 1SBOA	Xray (2.0)	17.1%	55 45

Tab. 3.5 Le strutture dei template identificate per la costruzione del modello del dominio STAS di Pendrina. Le strutture sono tutte relative a SpollAA di *Bacillus Sphaericus* (BACSH), *Bacillus Stearothermophilus* (BACS), *Bacillus subtilis* (BACSU) e *Thermotoga maritima* (THEMA). Per ciascuna struttura è riportato il codice PDB e il relativo indice di risoluzione in Angstrom. Vengono inoltre riportati i valori di similarità di sequenza con il dominio STAS di Pendrina e lo score di 3D-Jury ottenuto dalla ricerca del fold con Metaserver.

Come per il dominio transmembrana, si è voluto verificare l'attendibilità delle informazioni ottenute dalle banche dati con l'analisi della sequenza del dominio STAS utilizzando Metaserver per il riconoscimento di un fold proteico simile. Il Metaserver riconosce una possibile relazione tra la struttura secondaria del dominio STAS di Pendrina e tutte le strutture di SpollAA restituendo per ciascuna un punteggio ricavato da 3D-Jury. Tutti i punteggi ottenuti superano la soglia minima di affidabilità della predizione che è di 50 (Tab. 3.5).

La sovrapposizione delle strutture dei templati con CE ha permesso di individuare le regioni maggiormente conservate e quelle in cui esiste una certa variabilità strutturale tra le proteine ortologhe di SpollAA. Tra i diversi modelli, gli elementi di struttura secondaria mantengono la stessa disposizione anche se a livello dei β -strand 1 β e 3 β presentano lunghezze differenti. Le parti meno conservate si trovano in corrispondenza dei loop tra l'elica 1 α e il β -strand 3 β e tra le eliche 4 α e 5 α . In quest'ultimo loop alcune strutture presentano il β -strand 3 β . Inoltre, l'ultima elica al C-terminale del dominio STAS ha una disposizione spaziale leggermente diversa in tutte le strutture.

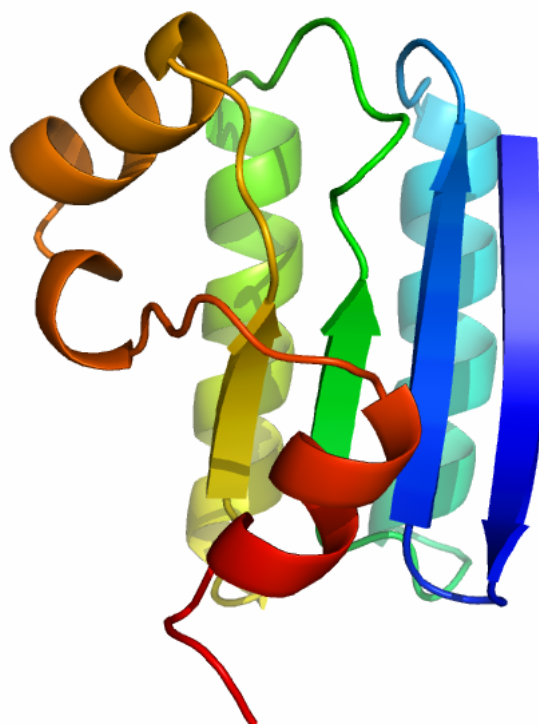


Fig. 3.10 Rappresentazione schematica del modello costruito per lo STAS domain di Pendrina. La struttura è colorata dal N- verso il C-terminale con gradazioni da blu a rosso.

La regione contenente il motivo conservato DXXG per il legame ai nucleotidi comprende il loop 3 β -2 α e parte dell'elica 2 α , regioni che alla sovrapposizione strutturale risultano molto conservati.

Le strutture sono state infine valutate con TAP da cui risulta che la struttura 1h4y è quella con migliore qualità. Data la similarità di sequenza, i risultati di TAP e le osservazioni ottenute dalla sovrapposizione strutturale è stato scelto come template la struttura di SpollAA di *Bacillus Sphaericus* (codice PDB: 1h4y).

Costruzione e descrizione del modello

L'allineamento migliore tra quelli ottenuti con GenAltAli è stato scelto in base ai valori di energia calcolati da FRST sui modelli costruiti con *Ali2Eval*. Dopo alcuni aggiustamenti manuali per posizionare il loop variabile di Pendrina nel modo migliore, l'allineamento è stato dato in input a Homer-M con il file PDB di 1h4y per costruire il modello. Il modello così costruito è stato valutato nelle sue componenti strutturali ed è stato possibile riconoscere alcune peculiari caratteristiche (Fig. 3.10).

Il calcolo della superficie elettrostatica con *APBS tools* di Pymol ha permesso di individuare due porzioni del modello con carica elettrostatica opposta (Fig. 3.11). In particolare la faccia della struttura costituita dalle eliche 2 α e 3 α e quella laterale tra l'elica 1 α e il β -strand 1 β hanno una superficie complessivamente positiva. La faccia dove si trova la substruttura a forma di mano e la faccia del loop conservato contenente il motivo DXXG mostrano invece una superficie elettrostatica negativa. Sulla struttura sono stati quindi mappati tutti i residui con carica positiva (Arg e Lys) e quelli con carica negativa (Glu e Asp). Le eliche 1 α e 2 α espongono sulla faccia esterna al dominio ad ogni giro di elica un residuo carico positivamente mentre l'elica 3 α della substruttura a forma di mano espone residui carichi negativamente. Queste eliche mostrano un pattern di residui idrofobici in ogni 4^a posizione interspersi da residui carichi e polari, caratteristica che rende le eliche anfipatiche con la faccia idrofobica rivolta verso il core del dominio e la faccia carica verso l'esterno. L'elica 3 α con la sua forma a mano nelle strutture di SpollAA viene mantenuta ancorata al resto del dominio grazie a due residui aromatici che si rivolgono verso il core, nel caso di Pendrina questi residui sono sostituiti da residui idrofobici ma nelle strutture che seguono o precedono l'elica altri residui idrofobici sono stati sostituiti con residui aromatici in modo da compensare la sostituzione e mantenere la stabilità strutturale. Ne dominio STAS di Pendrina, la substruttura a forma di mano viene ancorata al core dal residuo Phe-692, del β -strand 4 β esposto verso il core, e dal residuo Phe-708 dell'elica 4 α . A livello del 4B il residuo Tyr-691 tiene l'elica 4 α appoggiata al foglietto- β . Altri residui aromatici formano con i loro anelli un piano idrofobico sul lato del loop contenente il motivo conservato DXXG. I residui Phe-555, Tyr-556 e Phe-667 di questo piano sono posizioni conservate tra i trasportatori mentre il residuo Tyr-698 localizzato sull'elica 3 si riscontra nei trasportatori del solfato batterici.

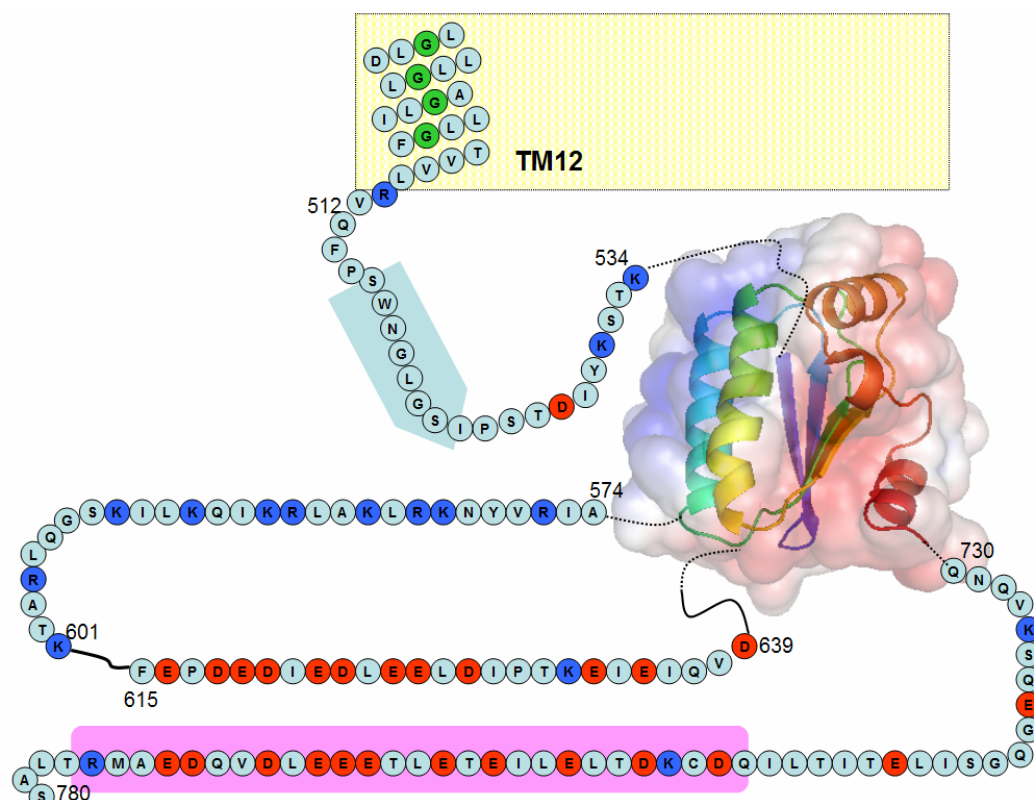


Fig. 3.11 Rappresentazione schematica della parte C-terminale di Pendrina. La struttura dello STAS domain è colorata dal N- verso il C-terminale con gradazioni da blu a rosso e la superficie elettrostatica semitrasparente. Le regioni loop sono indicate con la sequenza. Residui blu sono carichi positivi, rossi carichi negativi. Le glicine sono segnate in verde. La struttura coiled-coil è evidenziata in rosa, un beta strand in azzurro e la regione transmembrana in giallo.

Confronto con altre proteine del fold

Nella banca dati di Pfam il dominio STAS è classificato nel clan delle ATP-grasp un dominio ATP-grasp lega la molecola di ATP nella fessura formata dai due subdomini di cui è composto.

Dalla ricerca in SCOP risulta che esistono due superfamiglie di proteine che presentano un fold SpollAA-like. Una è quella nota degli antagonisti del fattore antisigma batterico e l'altra è quella del dominio CRAL/TRIO. Il dominio è detto appunto TRIO perché contiene in sé 3 domini uno dei quali ha un fold simile alle proteine ASA. La proteina Sec14 che contiene questo dominio catalizza lo scambio tra fosfatidil-inositolo (PI) e fosfatidil-colina tra il doppio strato lipidico. L'estremità C terminale di questa proteina contiene una tasca idrofobia che rappresenta il dominio di legame ai fosfolipidi. La sovrapposizione strutturale di alcune strutture contenenti il dominio CRAL/TRIO con la struttura di SpollAA ha

permesso di individuare le regioni conservate e quelle che differiscono maggiormente. Gli elementi di struttura secondaria mantengono la medesima posizione in entrambi i domini. Il loop tra 1 α e 3 β , che corrisponde al loop variabile dei trasportatori del solfato, la struttura viene mantenuta mentre grosse differenze si notano a livello del loop tra 2 β e 1 α che generalmente è un loop conservato per il dominio STAS. Un'altra differenza importante è a livello del loop tra 3 β e 2 α che in SpollAA è il loop conservato contenente il motivo di legame per i nucleotidi. Per le proteine Sec14 questo loop che è molto più esteso forma la tasca di legame per i lipidi.

Ricerca di motivi

La ricerca di motivi con ELM ha evidenziato la presenza di potenziali siti di legame per domini PDZ nell'estremità C-terminale, in posizione 743-746, 753-756 e 765-768. La ricerca di motivi è stata successivamente ripetuta utilizzando solo la sequenza del loop variabile di Pendrina. Sono stati così identificati nuovi siti di legame a domini PDZ in posizione 652-655 e nella regione centrale del loop in corrispondenza della sequenza in cui viene predetto disordine (posizioni 606-609, 618-621, 622-624, 624-627). Altri motivi interessanti sono in posizione 645-651, sito di legame per domini SH3; in posizione 593-596, sito di legame di glicosaminoglicani; 597-603, sito di fosforilazione di Kinasi cAMP-dipendenti. Sulla sequenza del dominio STAS di Pendrina sono stati inoltre ricercati manualmente motivi di legame al colesterolo $[L/V]-x(1-5)-Y-x(1-15)-[R/K]$. Questo pattern è stato riconosciuto nella sequenza 695-LQDYVIEK-702.

3.5 Analisi delle mutazioni

Il laboratorio MORL (Molecular Otolaryngology Research Laboratory) del Dipartimento di Otolaringologia dell'Università dell'Iowa cura un sito web in cui vengono raccolte tutte le mutazioni trovate in soggetti con Sindrome di Pendred o sordità di tipo DFNB4 (EVA). Integrando queste mutazioni con i dati più recenti di letteratura si possono contare 168 varianti di Pendrina, che comprendono 103 mutazioni missense, 8 mutazioni nonsense, 10 inserzioni, 20 piccole delezioni, 2 grosse delezioni, 19 siti di splicing, 1 duplicazione. 5 si possono considerare polimorfismi.

Tutte le mutazioni sono state mappate sulla sequenza di Pendrina ed è stata analizzata la loro distribuzione nelle tre diverse regioni studiate della proteina. Il 58% delle mutazioni trovate mappano nella regione transmembrana, il 21% nell'estremità C-terminale e il

16% nell'estremità N-terminale. Si nota inoltre che la frequenza dei siti mutati rispetto alle dimensioni delle regioni considerate è lievemente superiore nella regione transmembrana in cui si possono individuare due regioni con elevata incidenza di mutazioni. Una regione comprende le prime tre eliche transmembrana (residui 88-156) e l'altra è costituita dal nono segmento transmembrana e dal successivo loop extracellulare (residui 385-421). A livello del C-terminale le mutazioni si concentrano maggiormente nel tratto di residui che collegano la regione transmembrana con il dominio STAS e nella porzione C-terminale di quest'ultimo.

Il possibile effetto delle mutazioni sulla struttura e funzione di Pendrina è stato studiato in modo approfondito per le mutazioni dell'estremità C-terminale. In questo caso infatti è stato possibile mappare le mutazioni che comportano una sostituzione aminoacidica sul modello tridimensionale predetto del dominio STAS e valutare l'importanza della conservazione di ciascuna posizione.

Le mutazioni sono state suddivise in due classi, mutazioni inattivanti (Tab. 3.6) e quelle non-inattivanti (Tab. 3.7). Le mutazioni missense e le delezioni che portano alla perdita di un aminoacido sono state considerate non inattivanti mentre le mutazioni che creano un codone di stop o quelle che determinano un *frame shift*, insieme alle mutazioni che comportano un difetto di splicing costituiscono la classe delle mutazioni inattivanti. Per ogni mutazione è stata riportata la corrispondente variazione nucleotidica e aminoacidica, la sua posizione esonica e quella strutturale, e una possibile spiegazione sulla perdita di funzione della proteina che ne consegue. Per le delezioni, inserzioni o duplicazioni, laddove non era stato riportato in letteratura, è stato determinato il punto da cui inizia il *frame shift* e la posizione in cui viene a crearsi un codone di stop. Per alcune mutazioni è stato inoltre riportato il risultato ottenuto da test funzionali sulla funzione di Pendrina.

Per ogni mutazione è stato riportato il tipo di patologia riscontrata nei soggetti portatori della mutazione considerata, raccogliendo le informazioni disponibili in letteratura o fornite dal laboratorio di Diagnosi e Studio delle Malattie Rare della Clinica Pediatrica di Padova. Per 42 delle 58 mutazioni analizzate è stato possibile indicare il tipo di patologia a cui sono state associate. Risulta che alcune mutazioni sono state riscontrate in entrambi i tipi di patologia mentre altre sono associate a Sindrome di Pendred o sordità di tipo DFNB4.

3. Risultati

Esone/ Introne	Variazione nucleotidica	Mutazione e grado di conservazione	Posizione	Effetto	Attività di trasporto e targeting	Fenotipo	Referenze
13	1458_1459insT	1487YfsX526	Ultimo segmento di membrana	Probabile produzione di una proteina mancante dell'estremità carbossi-terminale dal residuo 526 che include il dominio STAS.	Completamente ritenuta in RE	DFNB4	Z. N. Brownstein et al., (2008)
13	c.1536-1538delAG	X524		L'introduzione di un codone di Stop potrebbe risultare nella produzione di una proteina tronca mancante del dominio STAS	Test funzionali su mutanti di pendrina mancante del dominio stas mostrano perdita di attività di trasporto	PS	Pryor et al., 2005
Introne 13	1544+9C>G; IVS13+9C>G	/	Regione che connette il dominio di membrana con il dominio STAS	Difetto di splicing non chiaro. I soggetti affetti presentavano altre due mutazioni patogenetiche	Nds	PS	Yong et al., 2001
14	1548insC	S517fsX526		L'inserzione della citosina comporta un frameshift a partire dal codone 517 con la creazione di un codone di Stop a 526. In questo modo si ha la perdita completa del dominio STAS	nds	nd	Park et al., 2003
14	1561_1571cttgggaatg gc	S523fsX548		Proteina tronca con perdita di gran parte del C terminale	↓↓↓ (I e CI)	PS	Fugazzola et al., 2007
14	1586delT	I529fsX546		Probabile produzione di una proteina tronca mancante di gran parte del dominio STAS	nds	DFNB4	Fitoz et al., 2007
Introne 14	1614+1G>A	Splice Donor	/	Difetto di splicing	nds	DFNB4	Blons et al., 2004
Introne 14	1615-7A>G	?	/	Difetto di splicing ?	nds	nd	?
Introne 14	1615-1G>A; IVS14-1G>A	Splice acceptor	/	Difetto di splicing	nds	DFNB4	Park et al., 2005
15	1652insT	S551fsX556	2° β-strand del dominio STAS	Probabile produzione di proteina tronca in corrispondenza del 2° β-strand del dominio STAS	nds	PS	Tsukamoto et al., 2003
Introne 15	1705+5G>A IVS15+5G>A	Splice donor	/	Difetto di splicing	nds	DFNB4	Park et al., 2005; Iwasaki et al., 2006 ; Yang et al., 2005

introne 15	IVS15-18T>A	/	/		Polimorfismo frequenza 10:100	nds	nds	DFNB4	Murgia A. personal communication
introne 16	IVS16-6G>A	splice donor?	/		In alcuni casi sostituzioni nucleotidiche in posizioni simili nelle regioni introniche comportano variazione nei livelli di trascritto normale o aberrante nei diversi tessuti.	nds	nds	DFNB4	Yang et al., 2005
17	1829C>T	S610X	Loop variabile: regione tra di cluster cariche positive e cluster di negative		Probabile produzione di proteina che si interrompe all'interno del loop variabile impedendo la formazione del dominio STAS.	nds	nds	DFNB4	Tsukamoto et al., 2003
17	1863delT	I621fsX634				nds	nds	DFNB4	Park et al., 2003
17	1873G>T	E625X	Loop variabile: cluster di negative		Probabile produzione di proteina che si interrompe all'interno del loop variabile impedendo la formazione del dominio STAS.	No N-Glicosilazione: ↓↓(Cl-/HCO ³⁻);	DFNB4	DFNB4	Yoon et al., 2008
17	1898delA	I634fsX634						PS	Van Hauwe et al., 1998
Introne 18	2089+1G>A; IVS18+1G>A	Splice acceptor	/		Difetto di splicing	nds	nds	DFNB4 PS	Blons et al., 2004; Murgia A. personal communication
19	2111insGCTGG	E704fsX722	3° α-elica del dominio STAS		Da 704 a 721 si ha un frame shift che comporta la traduzione di aminoacidi diversi da quelli della sequenza di Pendina. Si ha inoltre la produzione di proteina mancante degli ultimi 58 aminoacidi che comporta la perdita di parte dell'ultima elica del dominio STAS che appare conservata e coinvolta nel direccionamento del C terminale. In questa ultima regione è presente un motivo coiled-coil che sembra essere coinvolto nella stabilizzazione del dimero e nel suo cambiamento conformazionale.	nds	nds	DFNB4	Usami et al., 1999; Tsukamoto et al., 2003
19		F709fsX719	Loop 3a-5β		Come sopra.	nds	nds	DFNB4	Coyle et al., 1998;

3. Risultati

	2127delT								
19	2182insG	Y728X	(1)	C-terminale	Probabile produzione di proteina tronca mancante della parte C terminale contenente la struttura coiled-coil coinvolta nella dimerizzazione della proteina.	nds	PS	Fugazzola et al, 2000	
21	2343A>G	X781W(X823)			Porobabile produzione di una proteina con una estremità C terminale più lunga di 23 residui.	nds	PS	Lopez-Bigas et al., 2002	

Tab. 3.6 Mutazioni inattivanti. Legenda: * grado di conservazione secondo Conseq; ↓= 50 a 100% attività di trasporto rispetto a WT. ↓↓= 10 a 50% attività di trasporto rispetto a WT. ↓↓↓= <10% di WT; nds: non determinato sperimentalmente; Nd: non dichiarato. Localizzazione subcellulare: ER retention, membrane. PDT: perchhorate discharge test. **Fonte:** MORL <http://www.healthcare.iowa.edu/labs/pendredandbor/>

Esone/ Introne	Variazione nucleotidica	Mutazione e grado di conservazione	Posizione	Effetto	Attività di trasporto e targeting	Fenotipo	Referenze
13	1540C>A	Q514R (7*)	Regione che connette il dominio di membrana con il dominio STAS	Si suppone che mutazioni in questa zona interferiscano con il corretto posizionamento del dominio STAS	nds	PS	Pryor et al., 2005
13	1541A>G	Q514K (7*)			nds	PS, DFNB4	Pera et al., 2008
14	1579A>G	T527A (8*)			nds	nd	MORL personal communication
14	1588T>C	Y530H (7)	N terminale del β -strand 1 β del dominio STAS	Generalmente in questa posizione si trova un aa aromatico che contribuisce a formare il core idrofobico. Solo nei trasportatori batterici dove altri aa aromatici sono sostituiti da aa alifatici, si trova una His.	nds	PS DFNB4	Coyle et al., 1998; Campbell et al., 2001; Borck et al., 2003; Blons et al., 2004; Pryor et al., 2005
14	1589A>C	Y530S (7)			nds	DFNB4	Pryor et al., 2005
15	1655G>A	S552I (8)	Loop 2 β -1 α	L'amminoacido Ser 552 forma legami polari con la Ser 666 del loop adiacente β 3- α 2. La mutazione può quindi comportare una destrutturazione del dominio STAS.	nds	PS DFNB4	Blons et al., 2004
15	1666T>C	Y556H (9)	N-terminale elica 1 α	I residui aromatici in posizione 555, 556, 667, 698, nella struttura sono disposti con l'anello sullo stesso piano dal lato che si rivolge alla membrana plasmatica e che si suppone possa disporsi tra l'ambiente idrofobico polare e lo strato lipidico di membrana.	nds		Lopez-Bigas et al., 1999
15	1667A>G	Y556C (9)			Parziale distribuzione in membrana; $\uparrow\downarrow$	PS DFNB4	Coyle et al., 1998; Taylor et al., 2002; Van Hauwe et al., 1998
15	1694G>A	C565Y (5)	elica 1 α	Questo residuo è esposto verso l'esterno del dominio STAS, in altri trasportatori la posizione è occupata da aa diversi per caratteristiche chimico-fisiche. La sostituzione non dovrebbe creare grossi cambiamenti strutturali ma è possibile che questa alfa elica interagisca con il	nds	PS DFNB4	Van Hauwe et al., 1998 Tsukamoto et al., 2003; Pryor et al., 2005

3. Risultati

					loop variabile in determinate condizioni di forza ionica.					
16	1708G>A	V570I	(7)	C-terminale elica 1a	Posizione conservata perché contribuisce alla formazione del core idrofobico del dominio STAS. La sostituzione con una Ile non dovrebbe determinare una destrutturazione del dominio. La patogenicità di questa mutazione dovrebbe piuttosto essere spiegata da un difetto di splicing, dovuto alla variazione del sito donatore di splicing (primo nucleotide dell'esone 16)	nds	nds	nd	MORL Personal communication	
16	1717G>C	D573H	(3)	Loop variabile: Regione carica positivamente	Questa posizione, occupata prevalentemente da Asp o Glu, potrebbe influire sulla direzione in cui si estende il loop rispetto al dominio la cui conformazione si suppone importante per la rilevabilità di cambiamenti nella forza ionica intracellulare.	nds	nds	nd	MORL personal communication	
16	1743G>C	R581S	(6)		Riduce la carica positiva di questa regione	nds	DFNB4	Iwasaki et al., 2006		
16	1751C>A	A584E			L'introduzione di un residuo con carica negativa in questa regione potrebbe comportare un disequilibrio di cariche che impedisce l'impaccamento del loop variabile con il resto del dominio	nds	nd	MORL		
16	1790T>C	L597S			La composizione del loop variabile è caratteristica di ciascun paralog, questa posizione è una di quelle conservate tra pendrina e l'isoforma A3 (DRA) che sono vicine dal punto di vista filogenetico.	nds		DFNB4 PS	Campbell et al., 2001 Fugazzola et al., 2002 Blons at al., 2004 Pryor et al., 2005	
17	1826T>G	V609G	(2*)	Loop variabile: link tra la regione carica positiva e quella con residui negativi	Negli orologi di Pendrina questa posizione è occupata da una Gly, la presenza di una Val distingue la forma umana. Potrebbe trattarsi di un residuo responsabile dell'acquisizione di funzione propria di pendrina nell'uomo.	nds		DFNB4 (Monodalleica)	Pryor et al., 2005; MORL	

17	1870C>G	L624V	(2*)	Loop variabile: di cluster cariche negative	Negli ortologi di Pendrina questa posizione è occupata da una Pro, la presenza di una Leu distingue la forma umana. Potrebbe trattarsi di un residuo responsabile dell'acquisizione di funzione propria di pendrina nell'uomo	nds	nd	MORL
17	1958T>C	V653A	(1*)	Loop variabile	La sostituzione della Val con un residuo più piccolo come Ala potrebbe rendere più instabile il posizionamento del loop variabile.	↓ (l-e Ct)	DFNB4	Scotte et al., 2000
17	1970G>A	S657N	(7)	N-terminale del β-strand 3β	Ser 657 forma contatti polari con Lys del 2° β-strand e con la Gly del loop 1β-2β. Ha quindi un'importanza strutturale. Il gruppo polare OH è rivolto verso l'esterno del foglietto-β dove si appoggia la parte C-terminale del dominio STAS.	nds	DFNB4	Tsukamoto et al., 2003
17	1975G>C	V659L	(9)	β-strand 3β	La Val forma contatti polari con Lys e Ile del 2° β-strand, contribuendo a stabilizzare il foglietto-β. Negli ortologi di Pendrina questa posizione è spesso occupata da Val o Ile, ciò fa supporre che sia necessario un residuo in cui il Cβ sia legato a un H+ e un metile.	nds	DFNB4	Hu et al., 2007
17	1997C>T	S666F	(8)	Loop 3β-2α	Ser 666 è vicina a Ser 552 del loop 2β-1α quindi sembra avere un ruolo strutturale. Inoltre questa posizione rappresenta l'unico residuo polare in un segmento altamente idrofobico strutturalmente conservato anche nelle proteine ASA, dove esiste il motivo DXXG di fosforilazione.	nds	DFNB4 (bitali.)	Tsukamoto et al., 2003; Suzuki et al., 2007
17	2000T>G	F667C	(6)		I residui aromatici in posizione 555, 556, 667, 698, nella struttura sono disposti con l'anello sullo stesso piano dal lato che si rivolge alla membrana plasmatica e che si suppone possa disporsi tra l'ambiente idrofobico polare e lo strato lipidico di membrana.	nds	DFNB4 PS	Everett et al., 1997
17	2003T>C	L668P	(8)		In questa posizione per tutti i trasportatori analizzati vengono a trovarsi residui idrofobici. L'introduzione di una Pro a questo livello potrebbe comportare una diversa posizione del motivo conservato DVVG.	nds	nd	MORL

3. Risultati

17	2015G>A	G672E	(8)	2° α-elica del dominio STAS	La Gly è rivolta verso il core del dominio circondata da due Leu , una del loop precedente e una dell'elica. L'introduzione di un residuo con carica negativa nel core del dominio provoca una destrutturazione dello stesso.	Parziale distribuzione in membrana: ↓↓↓ (↓-)	DFNB4 PS	Coyle et al., 1998 Taylor et al., 2002
17	2027T>A	L676Q	(8)		Posizioni occupate generalmente da residui idrofobici. Leu e Phe si rivolgono verso il core del dominio STAS dove la sostituzione con un residuo polare, come Gln e Ser rende più instabile le interazioni tra gli elementi di struttura secondaria.	No N-glicosilazione; ↓↓↓ (Cl-/HCO ³⁻)	DFNB4	Park et al., 2003; Yoon et al., 2008
18	2048T>C	F683S	(7)			nds	nd	Prasad et al., 2004
18	2059G>T	D687Y	(6)	Loop 2α-4β	In questa posizione si ritrovano residui carichi o di piccole dimensioni. Asp è rivolto verso l'esterno del dominio dove si trovano diversi residui carichi. La sostituzione con Tyr causa una riduzione della carica elettrostatica negativa di quella regione, ritenuta importante per la funzione del dominio STAS.	nds	DFNB4	MORL
18	2080T>C	S694P	(7)	Loop 4β-3α	Ser prende contatti polari con la Lys della struttura a forma di mano del dominio STAS contribuendo all'impaccamento della struttura. La sostituzione con Pro comporta la perdita del contatto rendendo la struttura instabile.	nds	DFNB4 (PDI+)	Blons et al., 2004
19	2130C>T	D710D	(5)		Una same sense non ha nessun effetto a livello strutturale.	nds	nd	MORL
19	2139T>G	I713M	(3*)	Loop 5β-4α	La sostituzione aminoacidica non dovrebbe comportare un grosso cambiamento conformazionale, questa regione sembra la più flessibile e coinvolta nell'interazione con altro dominio STAS per formare il dimero	nds	nd	MORL
19	2162C>T	T721M	(9)	4° α-elica	La sostituzione porta alla perdita del gruppo polare che sembra fondamentale in questa posizione (occupata da Thr o Ser). Non è ben chiara la funzione di questa ultima parte del dominio STAS ma si pensa che possa avere un ruolo nella dinamica delle variazioni conformazionali del dominio.		DFNB4 PS	Usami et al., 1999; Tsukamoto et al., 2003; Blons et al., 2004 ; Murgia A personal communication
19	2168A>G	H723R	(7)		Questa posizione è molto conservata soprattutto tra vertebrati. L'anello di His è	Ritenzione in RE: NO N-	DFNB4	Yoon et al., 2008; Usami et al., 1999;

				rivolto verso il core del dominio, la sostituzione con un residuo carico comporta l'allontanamento dell'ultima elica dal resto del dominio STAS.	glicosilazione; $\downarrow\downarrow(\text{Cl}^-/\text{HCO}_3^-)$;	PS	Tsukamoto et al., 2003; Blons et al., 2004; Murgia A personal
19	2171A>G	D724G	(8)	La sostituzione con Gly, alifatico non polare, o Asn, polare non crico, risulta in un indebolimento della superficie elettrostatica che in questa faccia del dominio risulta essere complessivamente negativa. L'alternanza di superfici con carica opposta sembra avere un ruolo fondamentale nella funzione del dominio.	nds	nd	MORL
19	2170G>A	D724N	(8)	Polimorfismo	nds	DFNB4+ (PDT +)	Blons et al., 2004
19	2218G>A	G740S	(6*)	C terminale	Attività di trasporto di I- normale, in cellule TSA (Gillam et al., 2004)	Normal hearing soggetto con PS ma con due mutazioni in TPO (Thyroid peroxidase)	Pera et al., 2008
19	2219G>T	G740V	(6*)				
21	2326C>T	R776C	(6*)				

Tab. 3.7 Mutazioni non inattivanti. Legenda: * grado di conservazione secondo Conseq; $\downarrow$ = 50 a 100% attività di trasporto rispetto a WT. $\downarrow\downarrow$ = 10 a 50% attività di trasporto rispetto a WT. $\downarrow\downarrow\downarrow$ = <10% di WT; nds: non determinato sperimentalmente; Nd: non dichiarato. Localizzazione subcellulare: ER retention, membrane. PDT: perchlorate discharge test. **Fonte:** MORL <http://www.healthcare.uiowa.edu/labs/bendredandbor/>

4. Discussione

Un difetto di Pendrina porta a distinti effetti tessuto-specifici che si manifestano in due distinte patologie, la Sindrome di Pendred e sordità di tipo DFNB4. La Sindrome di Pendred è caratterizzata dalla combinazione di sordità neurosensoriale con un difetto tiroideo mentre nella sordità di tipo DFNB4 si osservano solamente gli effetti a livello dell'orecchio interno. Pendrina è un membro della famiglia degli scambiatori anionici SLC26, una famiglia di geni con rilevanza funzionale in diversi processi fisiologici e patofisiologici. La maggior parte dei membri SLC26 agiscono come scambiatori anionici e distinte isoforme differiscono significativamente in specificità anionica ma ad oggi le conoscenze sulle basi molecolari della loro diversità funzionale sono ancora limitate. L'applicazione di strumenti bioinformatici, in combinazione con le informazioni derivanti da esperimenti di laboratorio, ha permesso di proporre un modello strutturale di Pendrina che fornisce la base su cui costruire una ipotesi funzionale per Pendrina.

Punti luce e zone d'ombra sul modello transmembrana

L'immagine topologica di Pendrina, ottenuta dall'analisi in silico, risulta in una proteina transmembrana con l'estremità N-terminale di 82 residui e l'estremità C-terminale di 269 residui localizzate nel compartimento intracellulare, con una regione centrale transmembrana di 429 residui composta da un fascio di 12 α -eliche. Il C-terminale presenta una regione ordinata corrispondente al dominio STAS la cui struttura è stata modellata su quella sperimentalmente determinata di SpoIIAA di *Bacillus Sphaericus*.

Il consenso derivato dai risultati ottenuti con diversi metodi predittivi per la topologia di membrana ha permesso di definire i limiti dei segmenti transmembrana come indicato in Tab. 3.1. L'allineamento multiplo di sequenze della famiglia dei trasportatori del solfato ha inoltre permesso di verificare la conservazione di tali limiti tra sequenze di specie diverse e tra sequenze di paraloghi di Pendrina. Alcuni metodi utilizzati prevedono segmenti transmembrana che non comprendono quelli ottenuti dal consenso. Queste regioni, che sul modello definito da 12 eliche transmembrana si trovano in corrispondenza di alcuni loop, sono state analizzate nelle loro caratteristiche di composizione aminoacidica, di struttura secondaria, per la presenza del disordine e accessibilità al solvente. Le regioni comprese tra i residui 161-180 e 325-340, corrispondenti rispettivamente ai loop tra le eliche transmembrana TM3-TM4 e TM7-TM8, sono state escluse come possibili segmenti transmembrana. La

presenza di disordine e la mancanza di una struttura secondaria ad α -elica in queste regioni, in aggiunta alla presenza di diversi residui carichi e di residui di Prolina fa supporre che sia difficile per queste sequenze assumere una struttura ordinata che possa essere inglobata nel fascio di eliche transmembrana. All'interno dell'ambiente idrofobico di membrana infatti, la possibilità di formare strutture energeticamente stabili dipende dall'idrofobicità delle eliche transmembrana e dalla loro abilità nel formare interazioni stabili con altre eliche (Elofsson et al., 2007). A differenza delle regioni suddette, nella regione compresa tra i residui 235-270, corrispondente al loop tra le eliche TM5-TM6, SPRITZ predice una struttura secondaria ad α -elica e alcuni residui risultano non accessibili al solvente. La sequenza tra 250 e 263, in cui viene predetta l' α -elica, risulta conservata tra i trasportatori del solfato nelle caratteristiche di composizione aminoacidica. E' possibile infatti individuare un pattern di residui idrofobici in ogni 4° posizione interspersi da residui carichi o polari, caratteristica che rende le eliche anfipatiche. Il programma TOPMOD non rileva la presenza di regioni rientranti o eliche interfaccia in questa posizione ma dal grafico di ZPRED è possibile osservare che una breve sequenza nell'intorno di 260 ha un andamento parallelo all'asse della membrana. Questo potrebbe significare che questa regione rappresenti una potenziale elica interfaccia. I risultati (negativi) di TOPMOD si spiegano con la presenza di un solo residuo di Tirosina all'interno della sequenza considerata e con la stretta vicinanza, a questa regione, del limite in posizione 264 predetto per l'elica transmembrana TM6 da Prodiv-TMMHMM. Per la predizione di eliche interfaccia infatti TOPMOD si basa esclusivamente sul contenuto in residui aromatici (Y e W) e fa riferimento alla predizione dei segmenti transmembrana di Prodiv-TMHMM. Per il membro A5 (Prestina) della famiglia SLC26 è stato proposto un modello a 10 segmenti transmembrana in cui le eliche TM5 e TM6, conservate tra Prestina e Pendrina, sono state disegnate come regioni rientranti. Questo modello ha permesso agli autori di sostenere la localizzazione intracellulare del loop TM5-TM6 in cui sarebbe presente un potenziale sito di fosforilazione (S238 in Prestina) cGMP-dipendente (Deàk et al. 2005). In Pendrina questa posizione, non conservata tra i trasportatori del solfato, corrisponde al residuo N248 ed è localizzata in vicinanza della putativa elica interfaccia e i dati ottenuti da LIPS indicano una marcata differenza tra le eliche TM5 e TM6: l'elica TM5 è una delle eliche meno esposte ai lipidi mentre TM6 è una delle eliche maggiormente esposte. Inoltre dai dati di ZPRED sono le eliche TM4 e TM5 ad essere predette come regione rientrante. Si deve assumere che l'elica TM4 entri nella membrana dall'esterno della cellula. E' verosimile infatti fissare il loop TM3-TM4 come extracellulare basandosi sul fatto che in esso vengono

predetti siti di glicosilazione per Pendrina e in tutti i trasportatori del solfato. In questo modo si possono ipotizzare due situazioni. In un caso (Fig. 4.1 B), TM4 prosegue verso l'interno senza attraversare completamente la membrana e TM5 esce verso l'esterno. In questo modo, il loop TM4-TM5 rimane all'interno della membrana come dimostra la predizione di accessibilità in quella regione, e il loop TM5-TM6 si localizza nel compartimento extracellulare. Nell'altro caso (Fig. 4.1 C), l'elica TM4 entra ed esce dallo stesso lato della membrana formando una specie di U. Questo sarebbe possibile grazie alla presenza del motivo GXXXG sulla faccia maggiormente esposta ai lipidi che consente maggiore flessibilità all'elica. L'elica

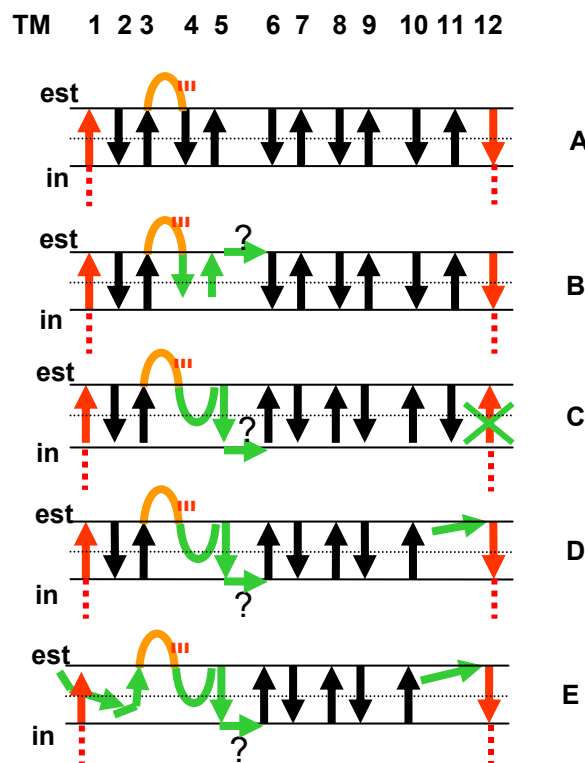


Fig. 4.1 Possibili configurazioni della topologia di membrana. Le frecce indicano segmenti transmembrana: in rosso sono le posizioni sperimentalmente determinate in Pendrina; in arancio il loop contenente il sito di glicosilazione; verde le posizioni dedotte dai dati ottenuti con i metodi di predizione per le regioni rientranti e eliche interfaccia. ?: elica interfaccia predetta all'interno del loop TM5-TM6. La figura (a) indica l'ipotesi di base, per la discussione delle altre figure vedi testo.

TM5 invece esce verso l'interno della cellula e il loop TM5-TM6 risulta nel compartimento intracellulare e tutte le altre eliche transmembrana cambiano direzione portando il C-terminale all'esterno della cellula, in contrasto con i dati sperimentali. Se però si considera un'altro dato proveniente dalla predizione di ZPRED, che descrive l'elica TM11 come elica interfaccia, è possibile considerare una terza situazione in cui le eliche TM6 - TM10

cambiano direzione; l'elica TM11 è parallela alla membrana disposta sul lato esterno a una distanza di 15Å dal centro della membrana e la TM12 si dirige verso l'interno della cellula posizionando il C-terminale in accordo con i dati sperimentali (Fig. 4.1 E). Quest'ultima configurazione fa sì che i loop TM6-TM7 e TM8-TM9, contenente il motivo di Saier, vengano posizionati all'esterno della membrana mentre in Prestina è stato dimostrato che questi loop risiedono nel compartimento intracellulare (Deak et al., 2005). La forte conservazione di queste regioni tra Pendrina e Prestina e altri trasportatori SLC26 permette di supporre che la struttura complessiva si mantenga simile e che un'inversione della disposizione delle eliche transmembrana sia un evento improbabile. Altre regioni di Pendrina presentano un andamento anomalo rispetto alle eliche che attraversano completamente la membrana. Nell'estremità C-terminale ZPRED predice un'elica transmembrana in corrispondenza del dominio STAS. E' possibile escludere che tale dato sia attendibile perché questa regione corrisponde all'incirca con il core idrofobico del dominio STAS e ZPRED si basa su una versione di ProdivTMHMM che in questo caso predice un'elica lontano dalla regione transmembrana. L'altro dato di ZPRED riguarda le prime tre eliche transmembrana in cui l'elica TM2 sembra allontanarsi di poco dal centro della membrana con un andamento pressoché parallelo all'asse della membrana e quindi perpendicolare alle altre eliche. E' interessante notare che l'elica TM2 e il successivo loop corrispondono al motivo conservato dei trasportatori SLC26. In questa regione era stata osservata una forte presenza di aminoacidi di piccole dimensioni, caratteristica di regioni rientranti, e di residui aromatici (3Tyr in TM2), indicativa di regioni interfaccia. Dai risultati di LIPS è possibile notare che le Tirosine sono disposte sulla faccia maggiormente esposta ai lipidi in contrasto con quanto ci si aspetta per una elica interfaccia. Inoltre se consideriamo l'elica TM2 come elica interfaccia la posizione del loop TM2-TM3 cadrebbe nel compartimento intracellulare. E' quindi verosimile pensare che l'elica TM2 rimanga in posizione centrale alla membrana e perpendicolare alle due eliche vicine con cui interagisce strettamente come dimostrato per SHST1 e Prestina (Shelden et al., 2001; Rajagopalan et al., 2006). La presenza di molti residui di piccole dimensioni correla con i dati ottenuti mediante mutagenesi sitospecifica sul motivo conservato di Prestina in cui l'analisi ET ha individuato due cluster di residui ravvicinati. L'interazione tra i cluster è permessa dalla giustapposizione di residui con determinate dimensioni e solo sostituzioni con aminoacidi di diversa grandezza hanno effetto sulla funzione della proteina. Inoltre la presenza di motivi GXXXG e GXXXA, rispettivamente nelle eliche TM1 e TM2, supporta l'ipotesi che ci sia un'interazione tra queste eliche. Secondo l'andamento dell'elica TM2 rilevata da

ZPRED e dai dati di LIPS è possibile stabilire che le due eliche interagiscono perpendicolarmente lungo le facce di interazione n°3 del modello a 7 facce.

La presenza di residui conservati di prolina rilevata a livello delle eliche TM2 e TM3 è un'altra importante caratteristica che aggiunge informazioni riguardo la struttura e la funzione di queste tre eliche. I residui di prolina creano dei punti deboli che potrebbero facilitare il cambiamento conformazionale di Pendrina dovuto al passaggio delle cariche, supportando l'ipotesi che le prime tre eliche e in particolare l'elica TM2 siano una componente del filtro di selettività della proteina.

L'evoluzione attraverso un loop

L'albero filogenetico, costruito con il multiallineamento delle sequenze del dominio STAS, ha permesso di confermare le relazioni intercorrenti tra i diversi paraloghi stabilite da altri studi filogenetici della famiglia SLC26. Questo dato fornisce un'informazione importante riguardo la coevoluzione del dominio transmembrana e del dominio STAS dei trasportatori SLC26 e permette di affermare che le differenze, presenti nelle sequenze del dominio STAS dei diversi paraloghi, contribuiscono in modo importante nel determinare la diversità funzionale dei trasportatori SLC26. All'interno della sequenza del dominio STAS è possibile infatti individuare una regione che mostra un'elevata variabilità tra i diversi paraloghi. Tale regione, corrispondente al loop $\alpha 1$ - $\beta 3$ precedentemente descritto come loop variabile del dominio STAS, varia tra i diversi paraloghi in composizione aminoacidica e in lunghezza. Lungo l'albero filogenetico è possibile osservare una crescita progressiva della lunghezza del loop soprattutto tra batteri, piante, funghi, nematodi e vertebrati. All'interno dei cluster di sequenze del dominio STAS dei trasportatori dei vertebrati invece avvengono eventi di inserzione e delezione che non rispettano più questa progressione. Appare comunque evidente che ciascun membro della famiglia SLC26 è caratterizzato da un ristretto *range* di lunghezza del loop variabile. Apparentemente non è possibile riscontrare nessuna zona di conservazione aminoacidica tra le sequenze all'interno del loop ma quello che accomuna questa regione è la presenza di cluster di residui carichi. Precedentemente è stato osservato che uno degli esoni in cui si riscontrano gran parte delle differenze in termini di inserzioni o delezioni tra le sequenze dei trasportatori SLC26 è l'esone 16 che in Pendrina codifica esattamente per la regione contenente il cluster di residui carichi positivamente (Okoruwa et al., 2008). All'interno del loop di Pendrina è presente anche un cluster di residui carichi negativamente all'estremità opposta del cluster positivo. Nei

vertebrati i trasportatori A4, A3, A5, A6, A1, A2 presentano due cluster di carica opposta all'interno del loop mentre i trasportatori con loop corto dei batteri mancano di questa caratteristica e quelli di *Arabidopsis thaliana* presentano un cluster di tre residui con carica negativa seguito da 2 residui con carica positiva. Nei membri A8 brevi cluster carichi con carica opposta sono interspersi lungo il loop. L'importanza di questi cluster è stata dimostrata in Prestina (A5) dove è stato supposto che gli anioni presenti nel compartimento intracellulare interagiscano con i cluster di residui carichi comportandosi da modulatori allosterici e favorendo le interazioni con altre proteine (Bai et al., 2006). In questo lavoro sono stati individuati diversi motivi di legame a domini noti all'interno del loop variabile di Pendrina tra i due cluster carichi. In particolare i motivi Lig_PDZ potrebbero essere il sito di legame a proteine PDZ adattatrici che mediano la formazione di multicomplessi tra Pendrina e altri trasportatori coinvolti nello stesso processo. A supporto di questa idea, esistono evidenze sperimentali che dimostrano l'esistenza di un'interazione tra i membri SLC26 (A3/DRA e A6) con CFTR in cui è coinvolto il dominio STAS.

Tracce di dimerizzazione

La struttura quaternaria oligomerica per i trasportatori SLC26A è supportata da diversi studi effettuati su Prestina. Più recentemente è stato dimostrato che lo stato oligomerico è una caratteristica conservata evolutivamente tra i trasportatori di questa famiglia. I risultati derivanti da diversi esperimenti risultano però contrastanti rispetto alla stechiometria delle subunità che formano gli oligomeri e al tipo di interazioni che permettono l'associazione tra i monomeri. Alcuni studi propongono una struttura quaternaria tetramerica in cui i monomeri sono tenuti insieme dai legami covalenti di ponti disolfuro, localizzati nel core della membrana plasmatica, mentre altri propongono una struttura quaternaria dimerica possibile grazie a interazioni non covalenti.

L'analisi in silico di Pendrina ha permesso di individuare alcuni elementi che sono implicati nel meccanismo di dimerizzazione. All'interno delle eliche transmembrana infatti sono stati identificati diversi motivi GXXXG che oltre a mediare l'interazione fra eliche vicine potrebbero essere punti di contatto con altre subunità di Pendrina. Il motivo GXXXGXXXGXXXG dell'elica TM12, in cui le glicine appaiono posizionate sulla faccia di probabile interazione con altre eliche, potrebbe mediare la dimerizzazione attraverso lo stesso motivo presente in monomeri diversi di Pendrina. Lo stesso motivo introduce inoltre un punto di instabilità che può promuovere una distorsione dell'elica. Il ruolo della glicina è stato studiato per proteine di cui è nota la struttura. Nel caso di KlrBac1.1,

un canale del potassio che funziona nello stato di tetrametro, le glicine, presenti nelle eliche, interne al poro del canale, di due subunità vicine, sono coinvolte nell'apertura del canale stesso (Kuo et al., 2003).

La struttura a coiled-coil predetta all'estremità C-terminale di Pendrina (CCMIII) è un altro elemento riconosciuto come motivo di dimerizzazione. Motivi coiled-coil mediano generalmente la formazione di dimeri o trimeri ma sono note anche interazioni coiled-coil che permettono di formare oligomeri di ordine elevato. L'analisi condotta su proteine ortologhe di Pendrina dimostra che il motivo CCMIII è conservato nei mammiferi e in alcuni paraloghi della famiglia SLC26 la cui bassa similarità di sequenza mostrata a livello dell'estremità C-terminale fa supporre che i motivi CCM rappresentino un tema strutturale e funzionale comune all'interno della famiglia. Il motivo CCMIII potrebbe essere introdotto o tolto in seguito a eventi di inserzione/delezione che sono particolarmente frequenti nelle regioni esoniche che codificano per le estremità N e C-terminale (Okoruwa et al., 2008). A supporto di questa idea sta il fatto che in Pendrina la sequenza in cui viene predetto l'elemento coiled-coil è codificata dall'esone 20 (aa 745-773), e i trasportatori SLC26 differiscono nel numero di esoni (vedi tab. 1.1).

E' possibile pensare quindi che la formazione di oligomeri non dipenda da un unico elemento ma che sia una combinazione di elementi che determina lo stato dimerico o tetramerico. Inoltre le interazioni coiled-coil in condizioni di pH acido vengono destabilizzate a causa della protonazione dei residui carichi, che determina la repulsione di residui in posizione e e g dell'heptad. I cambiamenti del microambiente possono in questo modo determinare variazioni conformazionali nella proteina (Kondabagil et al., 2006). Questo va a supporto degli esperimenti condotti su porzioni del dominio STAS in cui si dimostra che lo stato oligomerico dipende dal tipo e dalla concentrazione delle specie ioniche in soluzione.

La quarta dimensione

E' possibile ora proporre un modello complessivo di Pendrina che descrive la proteina non solo nelle sue componenti strutturali ma anche nei suoi possibili movimenti e interazioni. Questo modello prevede che Pendrina operi in una forma dimerica la cui stabilizzazione avviene mediante diversi punti di contatto. Uno di questi punti si trova all'interno della membrana ed è rappresentato dal motivo GXXXGXXXGXXXG dell'elica TM12 che permette interazioni tra i domini transmembrana di due subunità distinte. Gli altri punti di contatto prevedono l'avvicinamento del dominio STAS di due subunità distinte e la formazione della struttura coiled-coil tra

le estremità C-terminali. Due domini STAS sono avvicinati dalla parte dell'elica a forma di mano con una rotazione degli assi di 90° uno rispetto all'altro. In questo modo i loop variabili dei due domini si trovano in posizione opposta rivolti verso l'interno della cellula. Questa configurazione è stata supposta in base alla struttura cristallografica di un dimerico di SpoIIAA di *Thermotoga maritima*.

Capire come un segnale interno o esterno venga trasdotto attraverso la membrana aiuterebbe a comprendere la funzione di Pendrina. Nel modello proposto si suppone che un cambiamento nelle condizioni del microambiente determinino un cambiamento conformazionale, a livello del dominio citoplasmatico, che viene trasmesso al dominio transmembrana. Pendrina scambia anioni tra l'esterno e l'interno della cellula permettendo l'efflusso di Iodio o Bicarbonato verso l'esterno quando questi si accumulano all'interno della cellula. Un aumento degli anioni dovrebbe stabilizzare la struttura coiled-coil del C-terminale permettendo la formazione dello stato funzionale oligomerico di Pendrina. Un'altra conseguenza dell'aumento di cariche negative nel microambiente cellulare deve avere un qualche effetto sul dominio STAS che si presenta appunto come un elemento con facce di carica opposta. E' possibile che il loop variabile con i suoi cluster carichi sia conformato sul dominio STAS in modo da formare uno strato di cariche. L'interazione di ioni con la superficie dello STAS potrebbe destabilizzare o quasi scalciare i loop dalla loro posizione determinando un movimento rotatorio del dimerico che viene trasmesso al dominio transmembrana attraverso la regione di linking all'elica TM12. Lo spostamento della regione di linking potrebbe determinare una distorsione dell'elica TM12 a livello delle glicine e contribuire così all'apertura di un varco attraverso cui possono passare gli ioni. A contenere il flusso di ioni rimane il filtro di selettività costituito dalle tre eliche TM1-TM2-TM3.

All'interno della sequenza che forma la struttura coiled-coil e a livello del loop variabile sono presenti motivi Lig_PDZ che potrebbero mediare l'interazione con altri trasportatori coinvolti nello stesso processo e l'interazione dipenderebbe dallo stato in cui si trovano queste strutture.

Mutazioni

L'incidenza delle mutazioni osservata in particolari regioni della proteina permette di supporre che tali regioni siano particolarmente importanti per la funzione. A livello del dominio transmembrana infatti, le regioni dove localizzano le mutazioni corrispondono con i motivi conservati di Pendrina. Il ruolo fondamentale delle prime tre eliche è stato dimostrato da diversi studi e anche l'analisi in silico ha permesso di individuare peculiari caratteristiche di questa regione.

Rimane da chiarire il ruolo del motivo di Saier. A livello del C-terminale le mutazioni mostrano una maggiore incidenza nella regione di linking e nella parte C-terminale del dominio STAS. L'importanza di quest'ultima regione supporta l'ipotesi che rappresenti la superficie di interazione nel dimero di STAS. Sulla regione di linking non sono disponibili informazioni riguardo la sua struttura tridimensionale ma è possibile che l'elemento di struttura secondaria a β -strand predetto possa essere inglobato nel dominio STAS affiancandosi agli altri β -strand per formare il foglietto- β . Un'altra ipotesi è che il β -strand della regione di linking rappresenti il β -strand 1 β del fold SpollAA-like. I programmi di predizione infatti non permettono di predire il β -strand 1 β a livello della sequenza di Pendrina e di altri trasportatori del solfato. Gli elementi di struttura secondaria all'interno del dominio STAS sembrano essere codificati da specifici esoni, in particolare la regione di linking è codificata dall'esone 14 e il confine tra esone 14 e l'esone 15 cade vicino al β -strand 2 β . Potrebbe quindi essere stato inserito un loop più lungo tra i due β -strand. In questa regione sono presenti diversi residui aromatici che potrebbero avere un ruolo nell'accostare il dominio STAS alla membrana plasmatica. Le mutazioni trovate in questa regione infatti avvengono in posizioni occupate da aminoacidi polari o aromatici introducendo residui carichi o alifatici. Altre mutazioni avvengono nella faccia del dominio STAS che si rivolge alla membrana, in posizioni occupate da residui aromatici che formano un piano con i loro anelli. Il dominio STAS potrebbe essere addossato alla membrana attraverso interazioni non covalenti. Infatti, non esistono evidenze empiriche che dimostrino il C-terminale posizionato all'esterno della membrana. Esperimenti con tripsina o altri enzimi su Prestina indicano che le estremità della proteina sono protette (Bai et al., 2006).

Molte delle mutazioni missense studiate sul modello del dominio STAS colpiscono posizioni occupate da residui idrofobici, alcune delle quali costituiscono il core idrofobico e la cui sostituzione con residui carichi o polari crea una destrutturazione del dominio (V659L; L668P; G672E; L676Q). Tali mutazioni si possono definire "strutturali". Altre mutazioni che comportano la sostituzione aminoacidica di residui idrofobici (L597S; V609G; L624V; V653A) sono localizzate a livello del loop variabile, in corrispondenza dei siti di legame per domini PDZ predetti. Tali mutazioni si definiscono "funzionali". Altre mutazioni avvengono in posizioni occupate da residui carichi che contribuiscono alla superficie elettrostatica del dominio STAS (D687Y; H723R; D724G; D724N) o che costituiscono i cluster carichi del loop variabile (D573H; R581S). Data l'importanza delle cariche nel meccanismo di attivazione di Pendrina anche queste mutazioni si possono definire "funzionali".

Per 42 delle mutazioni studiate all'estremità C-terminale è stato possibile indicare il tipo di patologia presentata dai soggetti in cui sono state trovate allo scopo di delineare una potenziale correlazione genotipo-fenotipo. 14 mutazioni sono state associate a entrambe le patologie, 20 sono associate solo con sordità di tipo DFNB4 e 11 sono state trovate in soggetti con sindrome di Pendred. La scelta di suddividere le mutazioni in inattivanti e non inattivanti è stata applicata precedentemente per rilevare una correlazione tra il tipo di mutazione del gene GJB2 e la gravità del difetto uditivo, che varia da moderato a profondo in soggetti con mutazioni bialleliche. Questo studio ha permesso di individuare che certe classi di difetto uditivo si trovano in associazione con determinati genotipi quando le mutazioni vengono distinte in inattivanti e non inattivanti (Cryns et al., 2004). La suddivisione delle mutazioni trovate nell'estremità C-terminale di Pendrina in non inattivanti e inattivanti ha permesso di individuare una certa tendenza nel tipo di mutazioni che sono state associate a una sola delle due patologie. Delle 20 mutazioni trovate in soggetti con sordità DFNB4, 11 sono mutazioni inattivanti. Di queste, 2 sono varianti introniche che non avvengono nei siti di splicing canonici.

Delle 11 mutazioni associate solo con Sindrome di Pendred, 6 sono mutazioni inattivanti e 5 non inattivanti. Due soggetti portatori di variante missense non presentavano gozzo ma solo un test al perclorato positivo e un soggetto portatore di mutazione R776C presentava 2 mutazioni nel gene della Perossidasi Tiroidea (TPO) determinando una errata diagnosi di Sindrome di Pendred. Un soggetto portatore di IVS13+9C>G presenta altre due mutazioni patogenetiche. Risulta quindi che mutazioni inattivanti sono presenti in 5 di 7 (71%) pazienti con diagnosi clinica di Sindrome di Pendred e in 9 di 20 (45%) soggetti con sordità di tipo DFNB4. Questo dato conferma i risultati derivanti da studi funzionali in cui si dimostra che mutazioni che comportano una totale perdita della funzione di Pendrina sono maggiormente associate a Sindrome di Pendred.

5. Conclusioni

L'analisi condotta su Pendrina ha permesso di proporre un modello complessivo della proteina che conferma i dati sperimentali disponibili per alcuni trasportatori della famiglia SLC26 e nel contempo evidenzia caratteristiche peculiari di Pendrina.

Pendrina risulta essere costituita da un dominio centrale con 12 segmenti transmembrana e le estremità N- e C-terminali localizzate nel compartimento intracellulare. I metodi di predizione utilizzati per lo studio della regione transmembrana hanno permesso di individuare una potenziale elica interfaccia tra l'elica transmembrana TM5 e TM6 che non viene predetta dai comuni metodi per la predizione della topologia di membrana. Sono state inoltre individuate potenziali regioni rientranti a livello delle eliche transmembrana TM1-TM2-TM3 e TM4-TM5 e sono stati discussi diversi modelli topologici che divergono da quello tipico derivato da 12 tratti perpendicolari all'asse della membrana. Per le eliche TM1-TM2-TM3, date le loro caratteristiche di conservazione, di struttura 2,5D e di flessibilità, è stato proposto un ruolo nel determinare la specificità anionica dei trasportatori. La presenza del motivo GXXXGXXXGXXXG nell'elica TM12 e la predizione di una struttura coiled-coil all'estremità C-terminale supportano la struttura quaternaria dimerica, proposta per i trasportatori SLC26, e suggeriscono un meccanismo di attivazione di Pendrina che prevede il coinvolgimento del dominio STAS. In questo processo il loop variabile del dominio STAS riveste un ruolo fondamentale come sensore di anioni. Dallo studio filogenetico sulle sequenze del dominio STAS risulta inoltre che la lunghezza di questo loop è una caratteristica che differenzia i diversi paraloghi, dimostrando la sua importanza nel determinare la diversità funzionale tra i trasportatori SLC26.

Lo studio delle mutazioni riscontrate in Pendrina ha permesso di individuare potenziali regioni strutturalmente importanti per la funzione di Pendrina e di distinguere tra mutazioni strutturali e funzionali. La classificazione delle mutazioni in inattivanti e non inattivanti ha permesso di individuare una differenza tra le mutazioni riscontrate in soggetti con sindrome di Pendred e quelle associate a sordità di tipo DFNB4. Sembra infatti che la classe delle mutazioni inattivanti sia maggiormente rappresentata nella categoria di mutazioni associate solo a Sindrome di Pendred.

Lo studio eseguito rappresenta un esempio delle informazioni che si possono ottenere con i nuovi più sofisticati metodi di predizione per l'analisi di proteine transmembrana. Modelli topologici, residui potenzialmente esposti ai lipidi, potenziali regioni rientranti possono essere una guida per la pianificazione di esperimenti e per

l'interpretazione dei risultati. Indagini future potranno essere indirizzate nel verificare la localizzazione del loop TM5-TM6, contenente l'elica interfaccia predetta, che permetterà di definire la topologia di membrana. I risultati ottenuti da LIPS hanno permesso di individuare le potenziali facce di interazione fra le eliche di membrana. Studi di mutagenesi sito-specifica potranno indagare specifici siti di interazione basandosi sulle possibili combinazioni delle facce ottenute.

Lo studio delle mutazioni dovrebbe essere completato anche per la regione transmembrana e la possibilità di disporre di una risorsa accessibile in cui possano essere inseriti i dati relativi a ciascuna mutazione potrebbe aiutare a formulare una correlazione genotipo-fenotipo. Il sito MORL di cui ci si è serviti per ottenere la lista di mutazioni presenta scarse informazioni relative all'associazione con la patologia. Uno studio più accurato che consideri il genotipo potrebbe rilevare altre informazioni. Rimane infatti da spiegare la grossa percentuale di soggetti portatori di mutazioni monoallelica. A questo proposito sarà interessante lo studio dei siti di legame per domini noti, individuati in questo studio, che permetterà di conoscere la rete di interazioni in cui è coinvolta Pendrina e identificare possibili partner che collaborano nel determinare le patologie a cui è associata.

Bibliografia

- L. Adamian and J. Liang *Prediction of transmembrane helix orientation in polytopic membrane proteins*. BMC structural Biology, 6:13. (2006)
- L. Aravind and E. V. Koonin *The STAS domain – a link between anion transporter and antisigma-factor antagonists*. Current Biology, 10(2): 53-55. (2001)
- M. K. Byeon, A. Frankel, T. S. Papas, K. W. Henderson and C. W. Schweinfest. *Human DRA function as a sulfate transporter in Sf9 insect cell*. Protein expression and purification, 12: 67-74. (1998)
- K. Cryns, E. Orzan, A. Murgia, P. L. M. Huygen, F. Moreno, I. del Castillo, G. Parker Chamberlin, H. Azaiez, S. Prasad, R. A. Cucci, E. Leonardi, R. L. Snoeckx, P. J. Govaerts, P. H. Van de Heyning, C. M. Van de Heyning, R. J. H. Smith, G. Van Camp. *A genotype-phenotype correlation for GJB2 (connexin 26) deafness*. J Med Genet, 41:147–154. (2004)
- S. Detro-Dassen, M. Schanzler, H. Lauks, I. Martin, S. M. Berstenhorst, D. Nothmann, D. Torres-Salazar, P. Hidalgo, G. Schmalzing, C. Fahlke. *Conserved dimeric subunit stoichiometry of SLC26 multifunctional anion exchangers*. The American Society for Biochemistry and Molecular Biology. (2008)
- A. Elofsson and G. von Heijne. *Membrane protein structure: prediction versus reality*. Ann. Rev. Biochemistry, 76: 125-40. (2007)
- L. A. Everett, H. Morsli, DK Wu. *Expression pattern of the mouse ortholog of the Pendred's syndrome gene (PDS) suggest a key role for pendrin in the inner ear*. Proc Natl Acad Sci USA, 96: 9727-32. (1999)
- E. Granseth, H. Viklund and A. Elofsson. *ZPRED: Predicting the distance to the membrane center for residues in α -helical membrane proteins*. Bioinformatics, 22(14): 191-196. (2006)
- D. T. Jones. *Improving the accuracy of transmembrane protein topology prediction using evolutionary information*. Bioinformatics, 23(5): 538-544. (2007)
- K. R. Kondabagil and V. B. Rao. *A critical coil motif in the small terminase, gp16 from Bacteriophage T4: insight into DNA packaging initiation and assembly of packaging motor*. J. Mol. Biol., 358: 67-82. (2006)
- A. Kuo, J. M. Gulbis, J. F. Antcliff, T. Rahman, E. D. Lowe, J. Zimmer, J. Cuthbertson, F. M. Ashcroft, T. Ezaki, D. A. Doyle. *Crystal structure of the Potassium Channel KirBac1.1 in the closed state*. Science, 300: 1922-1926. (2003)

- G. Lasso, J. F. Antoniw and J. G. L. Mullins. *A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops*. *Bioinformatics*, 22(14): 290-297. (2006)
- D. Lee, O. Redfern and C. Orengo. *Predicting protein function from sequence and structure*. *Nature Reviews*, 8: 995-1005. (2007)
- W. Liu, E. Crocker, W. Zhang, J. I. Elliot, B. Luy, H. Li, S. Aimoto and S. O. Smith. *Structural role of glycine in amyloid fibril formed from transmembrane α -helices*. *Biochemistry*, 44: 3591-3597. (2005)
- K. Mio, Y. Kubo, T. Ogura, T. Yamamoto, F. Arisaka and C. Sato. *The motor protein prestin is a bullet-shaped molecule with inner cavities*. *The Journal of Biological Chemistry*, 283(2): 1137-1145. (2008)
- D. B. Mount and M. F. Romero. *The SLC26 family of multifunctional anion exchanger*. *European Journal Physiology* 447: 710-721 (2004)
- O. E. Okoruwa, M. D. Weston, D. C. Sanjeevi, A. R. Millemon, B. Fritsch, R. Hallworth and K. W. Beisel. *Evolutionary insights into the unique electromotility motor of mammalian outer hair cells*. *Evolution and development*. 10(3): 300-315. (2008)
- S. M. Owitt. *The role of cysteine residues in the sulphate transporter, SHST1: construction of a functional cysteine-less transporter*. *Biochimica e Biophysica Acta*, 95-100. (2005)
- H-J. Park, S. Shaukat, X-Z. Liu, S.H. Hahn, S. Naz, M. Ghosh, H-N. Kim, S-K Moon, S. Abe, K. Tukamoto, S. Riazuddin, M. Kabra, R. Erdenetungalag, J. Radnaabazar, S. Khan, A. Pandya, S-I Usami, W E Nance, E. R. Wilcox, S. Riazuddin, A. J. Griffith. *Origins and frequencies of SLC26A4 (PDS) mutations in east and south Asians: global implications for the epidemiology of deafness*. *Journal of Medical Genetics* 40: 242-248. (2003)
- E. Pasqualetto, A. Seydel, A. Pellini, R. Battistutta. *Expression, purification and characterisation of the C-terminal STAS domain of the SLC26 anion transporter prestin*. *Protein Expression and Purification*. (2008)
- A. Pera, M. Villamar, A Vinuela, M. Gandia, C. Meda, F. Moreno and C. Hernandez-Chico. *A mutational analysis of the SLC26A4 gene in Spanish hearing-impaired families provides new insights into the genetic causes of Pendred syndrome and DFNB4 hearing loss*. *European Journal of Human Genetics* 1-9. (2008)
- L. Rajagopalan, N. Patel, S. Madabushi, J. A. Goddard, V. Anjan, F. Lin, C. Shope, B. Farrel, O. Lichtarge, A. L. Davidson, W. E. Brownell and F. A. Pereira. *Essential Helix Interaction in the anion transporter domain of prestin revealed by evolutionary Trace analysis*. *The Journal of Neuroscience*, 26(49): 12727-12734. (2006)
- P. Rotman-Pikielny, K. Hirschberg, P. Maruvada, K. Suzuki, I. E. Royaux, E. D. Green, L. D. Kohn, J. Lippincott-Schwartz and P. M.

- Yen. *Retention of Pendrin in the endoplasmic reticulum is a major mechanism for Pendred syndrome*. Human Molecular Genetics, 11(21): 2625-2633. (2002)
- I. E. Royaux, K. Suzuki, A. Mori, R. Katoh, L. A. Everett, L. D. Kohn, and E. D. Green. *Pendrin, the Protein Encoded by the Pendred Syndrome Gene (PDS), Is an Apical Porter of Iodide in the Thyroid and Is Regulated by Thyroglobulin in FRTL-5 Cells*. Endocrinology, 141: 839-845. (2000)
- M. H. Saier, B. H. Eng, S. Fard, J. Garg, D. A. Haggerty, W. J. Hutchinson, D. L. Jack, E. C. Lai, H. J. Liu, D. P. Nusinew, A. M. Omar, S. S. Pao, T. T. Paulsen, J. A. Quan, M. Sllwinski, T-T Tseng, S. wachi, G. B. Young. *Phylogenetic characterization of novel transport protein families revealed by genome analyses*. Biochimica and Biophysica Acta, 1422: 1-56. (1999)
- M. Soleimani and J. Xu. *SLC26 Chloride/Base Exchangers in the Kidney in healt and disease*. Seminars in Nephrology, 26: 375-385. (2006)
- I. Sommer, S. Toppo, O. Sander, T. Lengauer and S.C.E. Tosatto. *Improving the quality of protein structure models by selecting from alignment alternatives*. BMC Bioinformatics, 7: 364. (2006)
- H. Suzuki, A. Oshima, K. Tsukamoto, S. Abe, K. Kumakawa, K. Nagai, H. Satoh, Y. Kanda, S. Iwasaki and S-I Usami. *Clinical characteristics and genotype-phenotype correlation of hearing loss patients with SLC26A4 mutations*. Acta Oto-Laryngologica, 1-6. (2007)
- S.C.E. Tosatto and S. Toppo *Large-Scale prediction of protein structure and function from sequence*. Current Pharmaceutical Design, 12: 2067-2086. (2006)
- H. Viklund, E. Granseth and A. Elofsson *Structural classification and prediction of Reentrant Region in α -helical transmembrane proteins: application to complete genomes*. J. Mol. Biol., 361: 591-603. (2006)
- T. Yang, H. Vidarsson, S. Rodrigo-Blomqvist, S. S. Rosengren S. Enerback and R. J. H. Smith *Transcriptional control of SLC26A4 is involved in Pendred syndrome and nonsyndromic enlargement of vestibular aqueduct (DFNB4)*. The American Journal of Human Genetics, 80: 1055-1063. (2007)
- S. Yohannan, S. Faham, D. Yang, J. P. Whitelegge and J. U. Bowie. *The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors*. PNAS 101(4): 959-963. (2004)
- J. S. Yoon, H-J Park, S-Y Yoo, W. Namkung, M. J. Jo, S. K. Koo, H-Y Park, W-S Lee, K. H. Kim and M. G. Lee. *Heterogeneity in the processing defect of SLC26A4 mutants*. BMJ. (2008)

- J. Zheng, G-G Du, K. Matsuda, A. Orem, S. Aguinaga, L. Deak, E. Navarrete, L. D. Madison and P. Dallos. *The C-terminus of prestin influences nonlinear capacitance and plasma membrane targeting*. Journal of Cell Science, 118: 2987-2996. (2005)