Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Biochimiche e Biotecnologiche

Ciclo XXVI

Settore Concorsuale di afferenza: 05/E1

Settore Scientifico disciplinare: BIO/10

METODI COMPUTAZIONALI PER L'ANNOTAZIONE DI GENOMI E PROTEOMI

Presentata da: **Valentina Indio**

Coordinatore Dottorato **Relatore**

Prof. Santi Mario Spampinato **Prof.ssa Rita Casadio**

Esame finale anno 2014

Sommario

ANNOTAZIONE DI GENOMI: **Analisi di dati Next Generation Sequencing nella ricerca sul cancro**

INTRODUZIONE

Il progresso tecnologico dell'ultimo decennio nel campo della biologia molecolare con la conseguente mole di dati prodotta, pone la comunità scientifica di fronte all'esigenza di dare un'interpretazione all'enormità di sequenze biologiche che a mano a mano vanno a costituire le banche dati, siano esse proteine o acidi nucleici. In questo contesto gli approcci multidisciplinari della biologia computazionale e della bioinformatica giocano un ruolo di primaria importanza, a cui si guarda con estrema attenzione nelle applicazioni mediche, delle scienze agrarie, veterinarie, microbiologiche.

Un nuovo livello di possibilità conoscitive è stato introdotto con l'avvento della tecnologia di sequenziamento massivo (NGS – acronimo da Next Generation Sequencing), per mezzo della quale è possibile ottenere interi genomi o trascrittomi in pochi giorni, una vera e propria rivoluzione se paragonata al tradizionale metodo di sequenziamento di tipo Sanger, con il quale in un singolo esperimento è possibile ottenere una sequenza di 50-1000 nucleotidi relativa ad una specifica porzione genomica nota.

Numerose sono le applicazioni delle tecnologie NGS: dal sequenziamento di nuovi genomi al risequenziamento di genomi già noti o parzialmente noti, da studi di carattere epigenetico alla metagenomica, dagli studi di espressione genica all'analisi del varioma di un qualsiasi individuo.

Tra le applicazioni del NGS più rilevanti ci sono senza dubbio quelle inerenti il campo oncologico. È ora possibile, infatti, effettuare la caratterizzazione genomica di tessuti tumorali la quale sta portando allo sviluppo di nuovi approcci diagnostici e terapeutici per il trattamento del cancro. Con l'analisi NGS è possibile individuare il set completo di variazioni che esistono nel genoma tumorale come varianti a singolo nucleotide, riarrangiamenti cromosomici, inserzioni e delezioni. Va però sottolineato che le variazioni trovate nei geni, di qualunque natura esse siano, vanno in ultima battuta osservate dal punto di vista degli effetti sulle molecole funzionalmente rilevanti, prime di tutte le proteine, in quanto esse sono le responsabili più dirette dei fenotipi alterati riscontrabili nella cellula tumorale.

L'expertise bioinformatica va quindi collocata sia a livello dell'analisi del dato prodotto per mezzo di NGS ma anche, e soprattutto, nelle fasi successive ove è necessario effettuare l'annotazione dei geni che sono contenuti nel genoma sequenziato e delle relative strutture proteiche che da esso sono espresse, o, come nel caso dello studio mutazionale, la valutazione dell'effetto della variazione genomica sul sistema biologico in cui essa viene riscontrata.

È in questo contesto che si colloca il lavoro di tesi che viene presentato nella trattazione: da un lato lo sviluppo di metodologie computazionali per l'annotazione di sequenze proteiche e dall'altro la messa a punto di una pipeline di analisi di dati prodotti con tecnologie NGS in applicazioni oncologiche avente come scopo finale quello della individuazione e caratterizzazione delle mutazioni genetiche tumorali a livello proteico.

ANNOTAZIONE DI PROTEOMI: **Predizione di Modificazioni Post-Traduzionali con Grammatical Restrained Hidden Conditional Random Fields**

BACKGROUND

Con il termine proteoma si fa riferimento all'insieme delle proteine di un organismo, ovvero alla totalità delle proteine codificate nel genoma, che in seguito ai processi di trascrizione, traduzione e maturazione espletano la loro specifica funzione all'interno di un sistema biologico[1].

Il proteoma è quindi l'insieme di tutti i possibili prodotti proteici espressi in una cellula, incluse tutte le isoforme e le modificazioni post-traduzionali.

La conoscenza sperimentale in ambito proteomico è in larga misura legata allo sviluppo delle tecniche cristallografiche e delle metodiche di sequenziamento diretto, che permettono rispettivamente di comprendere le proprietà delle proteine attraverso la visualizzazione delle strutture tridimensionali e di ottenere le relative sequenze amminoacidiche. La riuscita di tali esperimenti è, però, fortemente dipendente dal livello qualitativo della proteina di partenza. Generalmente, le proteine idrosolubili sono di più semplici da estrarre e purificare rispetto a quelle legate a matrici idrofobiche, quali le membrane cellulari[2]. Inoltre risulta difficoltoso l'ottenimento di proteine native, ovvero di proteine contenti i domini che vengono persi in seguito a maturazione.

È proprio a questo livello che si introducono le metodologie computazionali, che forniscono strumenti predittivi per l'annotazione di sequenze peptidiche per le quali la conoscenza sperimentale è ancora limitata.

In particolare il lavoro svolto nel corso del dottorato di ricerca, e presentato in questa tesi, riguarda lo sviluppo di due strumenti bioinformatici, entrambi volti alla predizione di modificazioni post-traduzionali: TPpred per la predizione di Targeting Peptide e SPpred per la predizione di Signal Peptide.

Targeting Peptide

Mitocondri e cloroplasti (nelle piante) sono organelli cellulari contenuti nella cellula eucariotica che svolgono processi metabolici e bioenergetici essenziali.
Secondo la teoria endosimbiotica, il DNA contenuto in questi organelli deriva da organismi batterici che sono stati incorporati nel citoplasma di cellule eucariotiche primordiali, tali organelli hanno trattenuto una porzione del genoma batterico originario che risulta codificante poche decine e poche centinaia di proteine (rispettivamente per mitocondri e cloroplasti).
Tuttavia, sia per mezzo di metodologie sperimentali che computazionali, si stima che migliaia di proteine siano localizzate all'interno di mitocondri e cloroplasti[3]. Da questa osservazione è quindi possibile dedurre che molte di queste proteine siano codificate dal genoma nucleare, sintetizzate nei ribosomi citoplasmatici e indirizzate verso gli organelli di destinazione per mezzo di segnali codificati all'interno della stessa sequenza peptidica.
Una delle più comuni tipologie di sequenza segnale consiste in un polipeptide N-terminale denominato *Targeting Peptide* (talvolta riportata come *Presenquence* nei mitocondri e *Transit Peptide* nei cloroplasti). Complessi proteici addetti al riconoscimento a alla traslocazione delle proteine dotate di Targeting Peptide guidano le catene peptidiche native verso il compartimento cellulare di destinazione dove successivamente il Targeting Peptide viene tagliato da specifici enzimi peptidasici. È stato stimato che in organismi pluricellulari circa il 10-25% dei geni codifica per proteine dotate di Targeting Peptide largamente eterogenei in termini di lunghezza (da 10 a 150 residui) e in termini di sequenza primaria ma con un organizzazione modulare comune[4].
Nella banca dati UniProtKB si contano circa 9200 sequenze dotate di Targeting Peptide delle quali solamente l'11% riporta un annotazione di tipo sperimentale. Quindi, data l'importanza biologica del meccanismo di targeting e la scarsità di conoscenza sperimentale in merito, sono stati introdotti metodi computazionali basati su differenti approcci di machine-learning. In questo lavoro di tesi sarà presentato TPpred, un nuovo metodo per la predizione della presenza del Targeting Peptide e della posizione del sito di taglio corrispondente basato si Grammatical Restrained Hidden Conditional Random Fields (GRHCRF).

Signal Peptide

Esiste in natura un altro tipo di sequenza segnale che comprende i residui N-terminali di un rilevante sottogruppo di sequenze proteiche sia eucariotiche che procariotiche e che, analogamente al Targeting Peptide, viene rimossa in fase post-traduzionale. Tale sequenza prende il nome di Signal Peptide e si differenzia dal Targeting Peptide in quanto presente in proteine destinate alla via secretoria e quindi

deputate ad attraversare o a costituire le membrane citoplasmatiche, dell'apparato di Golgi, del Reticolo Endoplasmatico, del nucleo e degli endosomi. È stato stimato che circa un terzo delle proteine codificate da un genoma contribuisce al processo di secrezione[5], è quindi verosimile che una percentuale di sequenze proteiche analoga contenga al suo interno il Signal Peptide addetto al trasporto della proteina attraverso la corretta membrana cellulare di destinazione.

Anche nel caso del Signal Peptide, la conoscenza sperimentale è piuttosto scarsa e metodologie computazionali in grado di predire la presenza e lunghezza di tali peptidi sono di grande interesse nell'ambito di studi proteomici. Presentiamo, in questa trattazione SPpred, un nuovo metodo per la predizione della presenza del Signal Peptide e della posizione del sito di taglio corrispondente, anch'esso basato su GRHCRF.

Stato dell'arte

Numerosi sono i metodi attualmente disponibili per la predizione del Targeting Peptide, basati su differenti approcci di machine-learning. I metodi meno recenti sono specifici per cloroplasti, come ad esempio ChloroP[6] e PCLR[7], o per mitocondri, come ad esempio MitoProt[8]. Altri tools più recenti integrano la predizione del Targeting Peptide con la predizione del Signal Peptide, sebbene, come vedremo, esistano numerose differenze tra i due tipi di segnale sia in termini di composizione che di lunghezza. Tra questi il metodo TargetP[9] che include ChloroP, iPSORT[10], Predotar[11] e PredSL[12]. Tutti i metodi citati predicono la presenta del Targeting Peptide, ma solamente MitoProt, TargetP e PredSL predicono la posizione del sito di taglio. In generale, tutti i predittori analizzano la regione N-terminale della proteina nativa, prendendo in considerazione i primi 40-100 residui a seconda del metodo. TargetP, PCLR, and Predotar si basano sulle Reti Neurali e l'input utilizzato include informazioni sulla composizione amminoacidica, l'idrofobicità e l'abbondanza dei residui dotati di carica. Il metodo iPSORT adotta un semplice algoritmo che genera una lista di decisione basata su regole riconducibili alla natura del dataset, mentre MitoProt definisce una funzione discriminativa basata su un insieme di 47 proprietà chimico-fisiche della sequenze in esame. Infine, PredSL combina Reti Neurali e Hidden Markov Models (HMM). Data la scarsità di conoscenza sperimentale, tutti i metodi sono stati addestrati con dataset che contengono anche sequenze con annotazioni elettroniche o ottenute utilizzando metodi comparativi.

Ancora più numerosi sono i metodi per la predizione di Signal Peptide anche in questo caso basati su differenti metodologie computazionali. Alcuni di questi tool sono stati implementati in modo specifico per effettuare la discriminazione tra Signal Peptide ed alpha-eliche transmembrana, tra questi troviamo SPOCTPUS[13] (metodo basato su reti neurali e HMM), Philius[14] (basato su Dynamic Bayesian Networks) e Phobius[15] (basato su HMM). Di quest'ultimo esiste una versione aggiornata denominata

POLYPhobius[16], sempre implementata con HMM, ma che incorpora un altro livello informativo basato sull'omologia di sequenza. Anche il metodo Signal-Blast[17] utilizza le informazioni ricavabili dalle sequenze omologhe a quella di interesse effettuando allineamenti di sequenza con il popolare tool Blastp (http://blast.ncbi.nlm.nih.gov). Metodi come Signal-CF[18] e PrediSi[19]per effettuare la predizione del Signal Peptide utilizzano le cosiddette "position weight matrix" mentre Signal-3L[20] fa uso di una tecnica di ensemble learning, infine SPEPlip[21] e SignalP4.1[22] utilizzano la tecnica di machine-learning basata sulle reti neurali. Tutti i metodi citati analizzano la porzione N-terminale della sequenza peptidica e, oltre a predire la presenza del Signal peptide, localizzano anche la posizione del putativo sito di taglio.

METODI

Machine-learning

La predizione del Targeting e del Signal Peptide può essere affrontata, in entrambi i casi, come un problema di labeling con forti condizioni grammaticali, in quanto si tratta di assegnare ai residui N-terminali di una sequenza peptidica l'etichetta "Y" (Signal o Targeting) o l'etichetta "N" (non Signal o non Targeting), dove i residui ai quali viene assegnata la label "Y" precedono sempre la regione etichettata come "N".

I Grammatical Restrained Hidden Conditional Random Fields (GRHCRF), sono stati introdotti recentemente come metodo di machine-learning per la risoluzione di problemi di labeling di sequenze biologiche[23]. I GRHCRF possono essere considerati come un'evoluzione dei Conditional Random Fields (CRF) per mezzo dei quali sono già stati approcciati numerosi problemi biologici: dalla predizione dei siti di interazione tra proteine[24], all'identificazione di geni all'interno di sequenze genomiche[2526], alla predizione di strutture secondarie di RNA[27] o di proteine transmembrana[28]. I GRHCRF, di più recente implementazione, sono stati già stati utilizzati con successo per la predizione della topologia di proteine transmembrana β-barrel nei procarioti[29], nella predizione di ponti disolfuro[3031]ed ora, nel presente lavoro di tesi, vengono proposti per la predizione del Targeting Peptide[32] e del Signal Peptide.

I GRHCRF offrono diversi vantaggi rispetto ai tradizionali metodi di machine-learning. Essi infatti, analogamente agli HMM, a partire dai quali sono stati introdotti, possono tenere conto della conoscenza pregressa sulle sequenze da trattare in quanto permettono di introdurre regole "grammaticali", ovvero vincoli che devono essere rispettati nella fase i labelling. Tuttavia, essendo modelli discriminativi, come gli Hidden Conditional Random Fields (HCRF), non richiedono le assunzioni di indipendenza che sono dovute negli approcci basati su HMM[33 23]. Inoltre, come le Reti Neurali, i GRHCRF possono analizzare input complessi ed eterogenei.

Risulta evidente, quindi, che il paradigma implementato nella presente trattazione include al suo interno le peculiarità dei principali metodi di machine-learning finora proposti per la risoluzione del problema di labelling di Targeting e Signal Peptide.

Dataset

I predittori TPpred e SPpred sono stati addestrati separatamente con due dataset differenti che vengono, quindi, illustrati in due sottosezioni distinte. Al contrario, la tipologia di input che viene

impiegata per addestrare i due predittori è la medesima e pertanto viene trattata univocamente al paragrafo "Input" di pagina 10.

Dataset di Targeting Peptide

Il set di sequenze utilizzate per addestrare il predittore di Targeting Peptide è stato selezionato dalla banca dati UniProtKB/SwissProt (release 11-2011) includendo le proteine complete più lunghe di 45 residui e con annotazione di esistenza come "evidence at the protein level". Da questo insieme è stato selezionato il set di esempi positivi (sequenze dotate di Targeting Peptide) cercando la parola chiave "TRANSIT PEPTIDE", la quale identifica all'interno della banca dati le sequenze peptidiche addette al direzionamento delle proteine verso un organello cellulare (***http://www.uniprot.org/keywords/KW-0809***). Sono state escluse da questo insieme le sequenze annotate come "by similarity", "probable" o "potential" e sono state incluse solo le sequenze di proteine mitocondriali o cloroplastiche con un sito di taglio noto. Le proteine senza la parola chiave "TRANSIT PEPTIDE", sono entrate a far parte del set di esempi negativi.

Adottando questi criteri di selezione abbiamo ottenuto 757 esempi positivi e 47363 esempi negativi. A partire da questo dataset è stata effettuata una procedura computazionale per generare un dataset non-ridondante. A questo proposito le sequenze sono state allineate a coppie con l'algoritmo BLASTP (considerando i primi 160 residui), raggruppate in cluster con livello di omologia superiore al 30% e con E-value inferiore a 10^{-3} (indice di significatività dell'allineamento). Il dataset non-ridondante finale è stato costruito selezionando in modo casuale una sequenza per ogni cluster. Esso consiste in 297 sequenze dotate di Targeting Peptide (113 di piante e 184 di organismi non vegetali) e 8010 senza Targeting Peptide (605 di piante e 7405 di organismi non vegetali).

Dato il notevole sbilanciamento numerico del dataset positivo rispetto a quello negativo è stata adottata una procedura di pre-training volta allo scopo di aumentare la capacità discriminativa del metodo predittivo.

Il GRHCRF è stato prima addestrato utilizzando solamente i 297 esempi positivi inibendo le transizioni dallo stato iniziale verso il sottomodello non-targeting (per i dettagli si veda la sezione *"Modello GRHCRF per TPpred"* della presente trattazione a pagina 16). Su questo modello addestrato sono stati poi predetti i 8010 esempi negativi forzando ad almeno 10 residui la predizione del Targeting Peptide. Quindi, per valutare l'affinità al modello degli esempi negativi, è stato calcolato uno score (S) come la somma della probabilità a posteriori della label "Y" per ogni posizione predetta come "Y" (da questo calcolo vengono escluse le prime 10 posizioni, in quanto risultano tutte con la medesima probabilità dato il vincolo del modello). Gli esempi negativi sono quindi stati ordinati in base a S (per definizione S assume valori compresi tra 0 e 1) e le sequenze con S>=0.95, vale a dir quelle più affini al modello, sono state incluse nel dataset di training come esempi negativi. Il dataset

di training finale consiste di 297 esempi positivi e 1245 esempi negativi corrispondenti a circa il 15% delle proteine comprese nel dataset negativo (Tabella 1).

Tabella 1. Dataset di addestramento di TPpred

Organismo	Senza TP (DB-)	Senza TP training(DB-)	TP Cloroplasti	TP Mitocondri	Con TP (DB+)
Piante	605 (86)	102	95 (12)	18 (0)	113 (12)
Non Piante	7,405 (1,081)	1143	-	184 (12)	184 (12)
Totale	8,010 (1,167)	1245	95 (12)	202 (12)	297 (24)

TP: Targeting Peptide. I numeri tra parentesi fanno riferimento alle sequenze dal dataset negativo che hanno una alpha-elica transmembrana annotata entro i primi 160 residui N-terminali.

Dataset di Signal Peptide

Il dataset di sequenze utilizzate per addestrare il predittore di Signal Peptide è stato selezionato distintamente per tre gruppi: batteri gram positivi (Gram+), batteri gram negativi (Gram-) ed Eucarioti. Nel dataset Gram+ sono stati inclusi le specie Actinobacteria e Firmicutes, ed escluse le specie Tenericutes in quanto mancanti di signal peptidasi di tipo I [34].

In maniera analoga alla metodologia descritta nel paragrafo precedente sono state incluse le sequenze di UniProtKB/SwissProt (release 01-2012), più lunghe di 45 residui, esclusi i frammenti e con evidenza sperimentale dell'esistenza della proteina. Per selezionare i dataset di esempi positivi è stata ricercata la parola chiave "SIGNAL", la quale identifica all'interno della banca dati le sequenze peptidiche presenti nella regione N-terminale di proteine destinate ad essere secrete o a costituire parte integrante delle membrane (http://www.uniprot.org/keywords/KW-0732). Come per le sequenze dotate di Targeting Peptide sono state escluse quelle con annotazione "by similarity", "probable" o "potential". Sequenze con Signal Peptide più corto di 15 o più lungo di 50 residui sono ste eliminate; inoltre, nei due gruppi di sequenze batteriche sono state esclusi i Signal Peptide annotati come Tat-type. Nei dataset di esempi negativi sono state incluse le sequenze senza la parola chiave "SIGNAL". Analogamente a quanto descritto per il Targeting Peptide, i tre dataset di Signal Peptide sono stati convertiti in insiemi non-ridondanti, con la sola differenza che in questo caso sono stati considerati i primi 60 residui nel calcolo dell'omologia di sequenza.

I dataset finali comprendono 1495 esempi positivi e 15714 esempi negativi per il gruppo Eucarioti, 194 esempi positivi e 922 esempi negativi per il gruppo Gram+ e 417 esempi positivi e 1741 esempi negativi per il gruppo Gram-. (Tabella 2).

Tabella 2. Dataset di addestramento di TPpred

Organismo	Con SP	Senza SP	Senza SP transmembrana
Eucarioti	1495	15714 (5000)	1775
Gram+	194	922	271
Gram-	417	1741	674

Si nota che per entrambi i gruppi di sequenze batteriche il set positivo e il set negativo sono numericamente bilanciati nella proporzione di ~ 1:4. Per fare in modo che tale proporzione sia rispettata anche nel dataset di sequenze Eucariotiche il set di 15714 esempi negativi è stato ridotto adottando una procedura di bilanciamento basata sull'idrofobicità della sequenza.
Per le sequenze sprovviste di Signal Peptide è stata quindi calcolato uno score di idrofobicità (HS) dei 60 residui N-terminali utilizzando la scala di Kyte-Dolittle utilizzando una finestra di 7 residui:

$$HS = \frac{\sum_{i=1}^{60} \frac{\sum_{j=i-3}^{i+3} Rj}{7}}{60}$$

Le 5000 sequenze con HS più elevato sono state incluse nel dataset di addestramento. Tuttavia, nella fase di test tutte le 15714 sequenze sono state considerate nella valutazione delle prestazioni del predittore.

Input

Nella fase di addestramento di entrambi i lavori proposti, sono stati considerati solamente i residui N-terminali di ciascuna sequenza: un segmento di 160 residui per il predittore di Targeting Peptide e di 60 residui per il predittore di Signal Peptide.
Ogni posizione di tali segmenti è stata codificata per mezzo di un vettore di 25 valori ciascuno dei quali descrive una caratteristica chimico-fisica del residuo corrispondente.
I 25 valori sono organizzati in quattro diversi moduli:

- I primi 20 valori descrivono la sequenza peptidica. Si tratta di valori di tipo binario (0,1) i quali assumono tutti il valore 0 ad eccezione di quello corrispondente al residuo presente nella posizione osservata (seq);
- Un valore che codifica il livello medio di idrofobicità (calcolato utilizzando la scala di Kyte-Dolittle) di un intorno di 7 residui centrato nel residuo corrispondente alla posizione osservata (kd);

- Due valori che rappresentano rispettivamente il numero di cariche positive e negative nell'intorno di 7 residui (ch);
- Due valori che descrivono il momento idrofobico calcolato considerando angoli di 100° e 160°, in modo tale da simulare rispettivamente la struttura di a-eliche e b-sheet. Per effettuare questi calcoli è stato impiegato il programma *hmoment* incluso in EMBOSS[35] considerando una finestra di 11 residui (hm).

Addestramento, validazione e test

Per addestrare e testare entrambi i predittori TPpred e SPpred è stata adottata una procedura 5-fold cross-validation. Il metodo consiste nel dividere in modo casuale i dataset non ridondante in 5 sottoinsiemi, tre dei quali vengono utilizzati per addestrare il modello, uno per la validazione e la scelta dei migliori parametri di input e l'ultimo per la fase di test e la valutazione delle performance. Per ciascuno dei 5 processi di training vengono valutate le performance per il relativo insieme di test. Nei due casi in cui parte degli esempi negativi sono stati eliminati dai set di training (dataset negativo di Targeting Peptide, e dataset negativo di eucarioti di Signal Peptide) ciascuna delle sequenze escluse viene predetta scegliendo casualmente uno dei 5 modelli addestrati. Tale scelta non va ad inficiare il risultato, le performance, infatti, non possono essere sovrastimate in quanto stiamo trattando sempre dataset non ridondanti le cui sequenze presentano livelli di omologia inferiori al 30%.

Calcolo delle performance

Per entrambi i predittori TPpred e SPpred sono stati utilizzati diversi indici per la valutazione delle performance di predizione. Si utilizzano le sigle TP e TN il numero dei veri positivi e dei veri negativi, e con FN e FP il numero di falsi negativi e di falsi positivi rispettivamente, con cui si indica:
Veri positivi: sequenze dotate di targeting o Signal Peptide che vengono predette come tali;
Veri negativi: sequenze sprovviste di targeting o di Signal Peptide che vengono predette come tali;
Falsi negativi: sequenze dotate di targeting o di Signal Peptide che vengono predette come sprovviste di Targeting o di Signal Peptide;
Falsi positivi: sequenze sprovviste di targeting o di Signal Peptide che vengono predette come aventi targeting o Signal Peptide.
I due indici che vengono calcolati per valutare le prestazioni globali dei predittori sono l'accuratezza globale (Acc) e il coefficiente di correlazione di Matthews (MCC), che sono definiti come segue:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FN) \cdot (TN + FP)}}$$

A differenza del Acc, il MCC è solo in minima parte influenzato dallo sbilanciamento tra esempi negativi e positivi all'interno del dataset e viene quindi considerato come il più attendibile e significativo degli indici di performance globali.
Oltre ad Acc e MCC vengono calcolati altri indici di performance per entrambe le classi predittive (targeting, non-targeting o signal, non-signal) che sono: la sensitività (Sn), la specificità (Sp) e la percentuale di falsi positivi (FPR). Tali indici sono calcolati come segue:

$$Sn(c) = \frac{TP}{TP + FN}$$

$$Sp(c) = \frac{TP}{TP + FP}$$

$$FPR(c) = \frac{FP}{TN + FP}$$

Dove c è la classe presa in considerazione. Ulteriori dettagli sul significato dei seguenti indici si rimanda a Baldi et al. (2000).

RISULTATI

Risultati TPpred

Caratteristiche dei Targeting Peptide

Lunghezza: La lunghezza dei Targeting Peptide mitocondriali e cloroplastici varia da 10 a 150 residui (Figura 1) con medie rispettive di 35 e 59 residui. I Targeting Peptide mitocondriali risultano quindi più corti rispetto quelli cloroplastici anche se la dispersione intorno alle lunghezze medie è in entrambi i casi molto elevata, rispettivamente di circa 16 e 35 residui. Nonostante queste rilevanti differenze in termini di lunghezza, è stato deciso di non separare il dataset in sequenze mitocondriali e cloroplastiche per due ragioni: la scarsità di esempi con annotazione sperimentale di Targeting Peptide e la possibilità che lo stesso Targeting Peptide possa mediare la traslocazione della proteina sia verso il mitocondrio che verso il cloroplasto (Carrie et al., 2009).

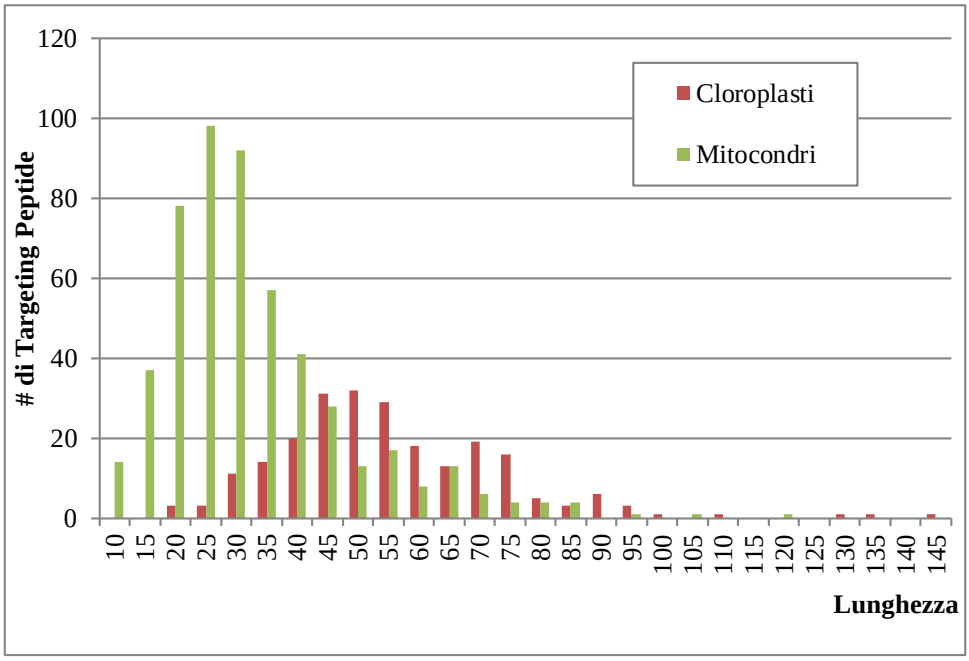


Figura 1. Distribuzione della lunghezza del Targeting Peptide del dataset di esempi positivi

Composizione dei residui: La composizione dei residui dei Targeting Peptide è mostrata nella Figura 2 nella quale vengono confrontate le frequenze amminoacidiche dei Targeting Peptide con quelle delle

intere sequenze incluse nel dataset. Data la notevole quantità di informazioni mostrate nella Figura 2, è stato deciso di non mostrare la deviazione standard che è stata stimata essere circa il 20%. Le composizioni delle intere sequenze mitocondriali e plastidiche (rappresentate nella Figura 2 rispettivamente in azzurro e giallo) non mostrano particolari differenze con le proteine del dataset negativo (in nero). Al contrario, i Targeting Peptide mostrano una composizione fortemente caratteristica. Sia mitocondri che cloroplasti (verde e rosso rispettivamente) presentano Targeting Peptide ricchi di alanina (A) e nettamente scarsi di residui carichi negativamente (D, E), Tirosina (Y) e Isoleucina (I). Altre differenze rilevanti possono essere evidenziate se si esaminano separatamente i Targeting Peptide mitocondriali e cloroplastici: i primi, infatti, risultano più ricchi di Metionina (M), Leucina (L) e Arginina (R), mentre il residuo Lisina (K) è decisamente sotto rappresentato; i Targeting Peptide plastitici, invece, contengono più Serina (S) e Prolina (P) a discapito della Glicina (G) che è meno frequente. La composizione dai Targeting Peptide studiata e presentata nel nostro dataset è in accordo con studi precedenti (Texeira and Glaser, 2012).

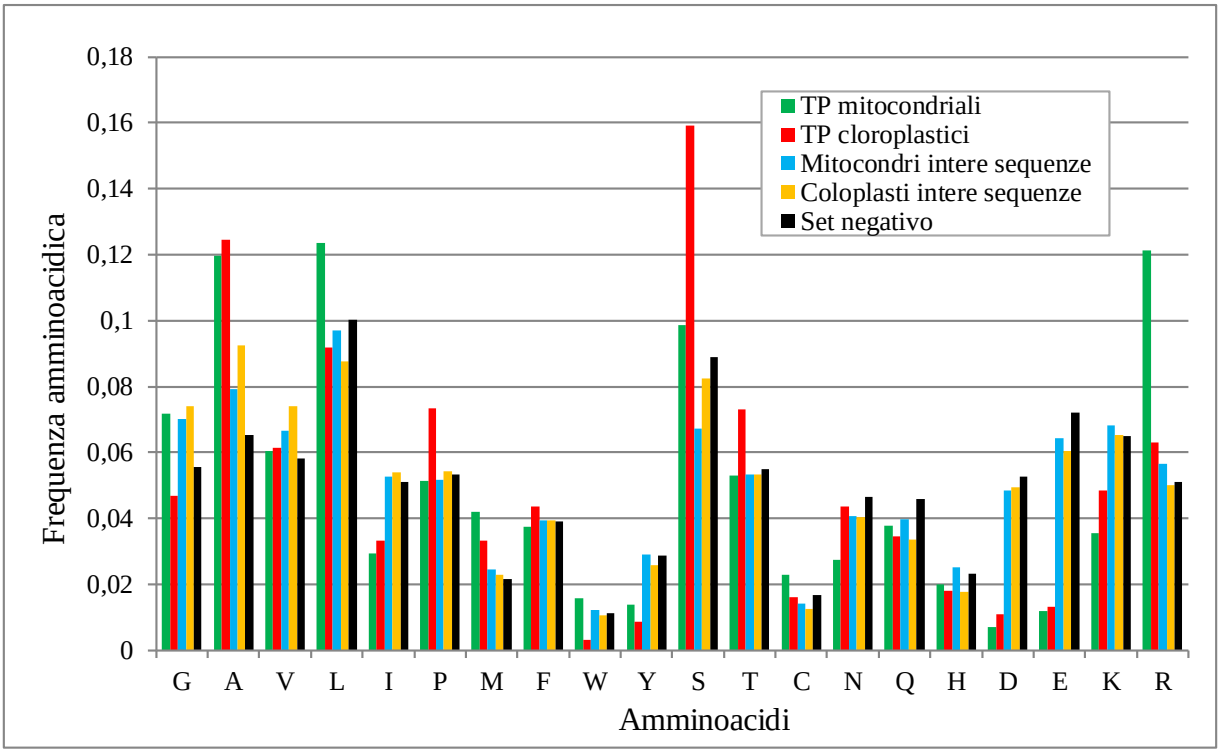


Figura 2. Composizione amminoacidica dei Targeting Peptide e delle intere sequenze incluse nel dataset di training di TPpred. I valori solo calcolati su 202 sequenze mitocondriali, 95 cloroplastiche e 8010 proteine senza targeting peptide (set negativo). L'errore standard, stima della dispersione intorno ai valori medi, è di circa il 20% e non è rappresentato in figura.

Sito di taglio: L'informazione composizionale descritta nel paragrafo precedente risulta essere più forte e maggiormente significativa nelle regioni vicine al sito di taglio, dal momento che in esse risiede il segnale di riconoscimento per i siti attivi del complesso peptidasico. Le 297 proteine del dataset positivo sono state allineate considerando gli 8 residui a monte e a valle del sito di taglio e il profilo risultante è visualizzabile nel "sequence logo" in Figura 3. Tale rappresentazione mette in evidenza il livello di conservazione di ogni residuo presente in una certa posizione, l'altezza di ciascuna lettera è proporzionale alla frequenza del residuo amminoacidico corrispondente in un intervallo che va da 0 a $\log_2 20 = 4.3$. Dalla Figura 3 è evidente che le posizioni intorno al sito di taglio sono più conservate rispetto alle altre anche se il contenuto informativo appare comunque moderato. Le posizioni maggiormente conservate risultano essere la -1 e -2 a valle e la +1 e +2 a monte del sito di taglio.

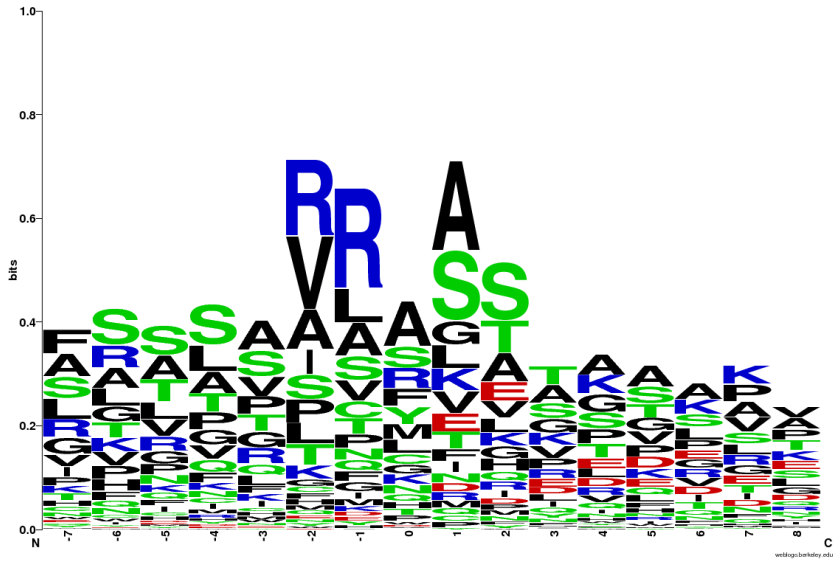


Figura 3. Sequence logo delle posizioni intorno al sito di taglio dei targeting peptide. Il sequence logo è stato calcolato per mezzo dell'applicazione online WebLogo (http://weblogo.berkeley.edu/). La posizione "1" indica il primo residuo della sequenza matura. L'altezza delle lettere è proporzionale alla frequenza dell'amminoacido nella posizione corrispondente. I codice colore è organizzato per distinguere i residui idrofobici (nero), polari (verdi), carichi positivamente (blu) e carichi negativamente (rosso).

Modello GRHCRF per TPpred

Le caratteristiche discusse nei paragrafi precedenti sono state incorporate in un modello, ovvero nell'automa illustrato nella Figura 4. Tale modello è stato adottato per generare il labelling dei 160 residui N-terminali delle sequenze sottomesse.

Lo stato "begin" è lo stato iniziale dal quale sono possibili due transizioni alternative verso i due sottomodelli i quali descrivono rispettivamente le sequenze dotate di Targeting Peptide (modello targeting, in azzurro) e sprovviste di Targeting Peptide (modello non-targeting, in arancio). Il modello consiste in:

- 45 stati con label "Y" rappresentati con quadrati e volti a modellare il Targeting Peptide;
- 31 stati con label "N" rappresentati con cerchi e volti a modellare le sequenze sprovviste di Targeting Peptide e i residui a valle dei targeting peptide.

A sua volta il sottomodello targeting è composto da quattro moduli:

- Una regione N-terminale;
- Due domini intermedi utili per descrivere la struttura modulare dei Targeting Peptide
- La regione del sito di taglio con tre residui a monte e quattro a valle obbligati;
- La regione a valle del sito di taglio.

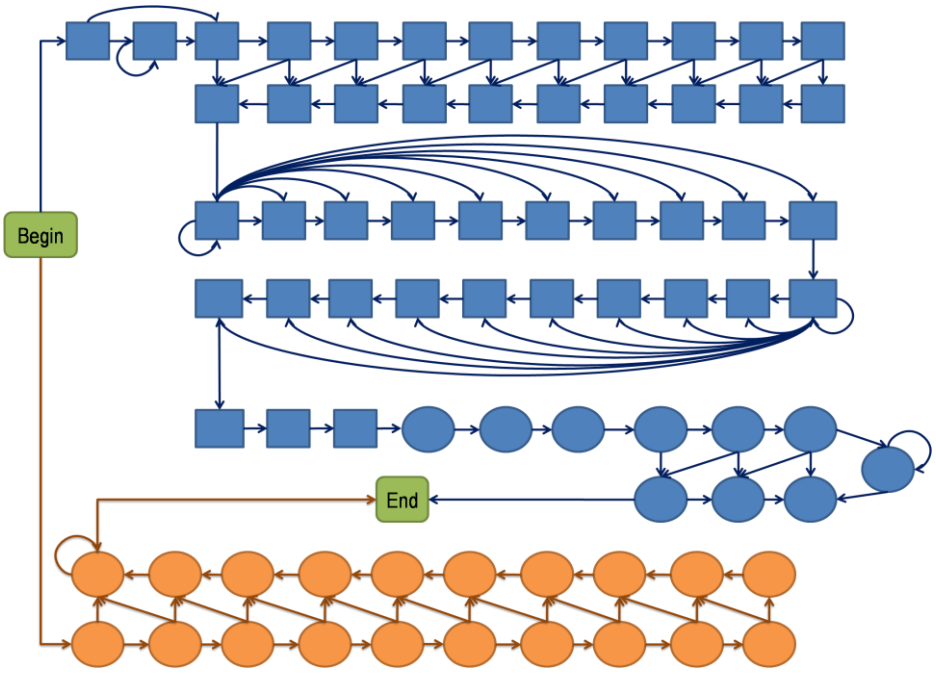


Figura 4. Modello predittivo di Targeting Peptide. I quadrati rappresentano gli stati con label "Y" (targeting peptide) mentre gli cerchi rappresentano gli stati con label "N" (non-targeting). Gli stati blu rappresentano il sottomodello targeting mentre gli stati arancioni il sottomodello non-targeting.

Secondo questa topologia, ciascuna sequenza analizzata dal sottomodello targeting deve contenere almeno 10 residui etichettati come "Y" (targeting) e 5 come "N" (non-targeting). Ne consegue, in maniera congrua con l'analisi effettuata sul dataset (Figura 1) che un Targeting Peptide predetto dal modello descritto potrà avere una lunghezza che va da 10 a 155 residui (dal momento che viene predetta una sequenza N-terminale di 160 residui).

Il sottomodello non-targeting è molto più generico ed è stato ideato per catturare la grande variabilità delle proteine incluse nel dataset negativo, pertanto non presenta una architettura modulare bensì una struttura quanto più possibile priva di vincoli.

Predizione con differenti input

Il GRHCRF è stato addestrato adottando la strategia descritta nel paragrafo "Addestramento, validazione e test" di pagina 11 per mezzo di una procedura 5-fold cross validation: tre set sono stati impiegati per il training, uno per scegliere automaticamente i parametri del modello (set di validazione) e uno per calcolare gli indici di performance (set di test). Dal momento che il dataset è stato globalmente ridotto per omologia (si veda il paragrafo "Dataset di targeting peptide") tale procedura assicura risultati attendibili ed evita il problema dell'overfitting.

La Tabella 3 mostra le prestazioni del predittore sul set completo di sequenze comprese nel dataset (297 esempi positivi e 8010 esempi negativi) a seconda della tipo input adottato. Utilizzando come input solamente la singola sequenza si ottiene un'accuratezza globale (Acc) del 95% e un Matthews Correlation Coefficient (MCC) di 0.50. Aggiungendo a mano a mano le differenti caratteristiche descritte nel paragrafo "Input" di pagina 10 le prestazioni del predittore aumentano fino a raggiungere i massimi valori quando l'input è composto da sequenza, idrofobicità, carica e momento idrofobico. Le performance finali di TPpred ammontano al 96% di Acc e a 0.58 di MCC.

Tabella 3. Performance di GRHCRF con diverse codifiche di input per la predizione del Targeting Peptide.

Input	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(-) (%)	Sp(-) (%)	FPR (%)
seq	95	0.5	69	40	96	99	3.8
seq+kd	95	0.54	73	42	96	99	3.7
seq+kd+ch	96	0.56	75	46	97	99	3.3
seq+kd+ch+hm (TPpred)	96	0.58	75	48	97	99	3

seq: sequenza; kd: livello di idrofobicità calcolato con la scala di Kyte-Doolittle; ch: carica ; hm: momento idrofobico. La valutazione riportata in tabella comprende 297 esempi positivi e 8010 esempi negativi. Gli indici di performance sono calcolati con una procedura di 5-fold cross-validation.

Confronto con i predittori allo stato dell'arte

La Tabella 4 mostra le performance predittive di TPpred su un dataset non ridondante a confronto con quelle ottenute dagli altri metodi disponibili attualmente per predire Targeting Peptide mitocondriali e cloro plastici. TPpred risulta essere il miglior predittore.
La dichiarazione di TPpred come miglior predittore assume ancora maggiore forza se si considera che il nostro predittore è l'unico testato in cross-validation mentre, per tutti glia altri predittori testati, la sovrapposizione tra le sequenze di training e le sequenze di test comporta che le prestazioni di tali tool siano sovrastimate.

Tabella 4. Risultati ottenuti con TPpred a confronto con i predittori allo stato dell'arte non specifici per mitocondri o cloroplasti

METODO	Acc %	MCC	Sn(+) %	Sp(+) %	Sn(-) %	Sp(-) %	FPR %	FPR TM[#] %
TPpred*	96	0.58	75	48	97	99	3	2.5
TargetP°	89	0.44	93	24	89	100	10.9	11.1
Predotar°	92	0.49	90	30	92	100	7.7	8.8
iPSORT°	89	0.38	79	22	90	99	10.2	12.4
PredSL°	91	0.45	88	26	91	100	9.4	12.5

Gli indici di performance sono descritti al paragrafo "Calcolo delle performance" a pagina 11. La valutazione riportata in tabella comprende 297 esempi positivi e 8010 esempi negativi. * Le sequenze incluse nel set di training sono predette in cross-validation, le sequenze del dataset negativo non incluse nel dataset di training sono predette selezionando in modo casuale uno dei 5 GRHCRF addestrati. ° Il dataset di confronto si sovrappone con il dataset di training. # Il tasso di falsi positivi (FPR) è calcolato sul set di esempi negativi che sono dotati di una alpha-elica transmembrana nei 160 residui N-terminali (Tabella 1, valori tra parentesi).

Il maggior svantaggio dei predittori esistenti è la bassa specificità per la classe positiva (SP+), fatto probabilmente dovuto allo sbilanciamento nei dataset utilizzati per l'addestramento. TPpred, in termini di specificità è nettamente superiore al migliore dei predittori testati (48% di TPpred contro 30% di Predotar) anche se questo si ripercuote in una sensitività più bassa (Sn+ = 75%). Un altro punto a favore di TPpred è il basso tasso di falsi positivi (FPR=3%) che risulta essere meno della metà rispetto tutti gli altri. Lo stesso trend è confermato quando confrontiamo TPpred con predittori specifici di Targeting Peptide mitocondriali (MitoProt) o cloroplastici (PCLR): anche in questo caso il nostro tool presenta i più alti Acc e MCC (Tabelle 5a e 5b).

Abbiamo anche testato il FPR dei vari predittori (compreso il nostro) su un sottoinsieme di esempi negativi che presentano un'elica transmembrana entro i 160 residui N-terminali (ultima colonna delle

Tabelle 4 e 5). Si nota che meno del 2.5% delle sequenze incluse in questo sottoinsieme vengono predette come a aventi Targeting Peptide da TPpred e, anche in questo caso, questo è il miglior risultato ottenuto tra tutti i metodi testati. Questa osservazione fa sì che TPpred possa essere considerato un buon metodo per analizzare proteine di membrana con topologia all-alpha.

Tabella 5a. Risultati ottenuti con TPpred a confronto con MitoProt, predittore specifico per sequenze mitocondriali.

Metodo	Acc %	MCC	Sn(+) %	Sp(+) %	Sn(-) %	Sp(-) %	FPR %	FPR TM# %
TPpred*	96	0.52	74	39	97	100	3	2.5
MitoProt°	77	0.23	89	9	76	100	22	29.9

Tabella 5b. Risultati ottenuti con TPpred a confronto con PCLR, predittore specifico per sequenze cloroplastiche.

Metodo	Acc %	MCC	Sn(+) %	Sp(+) %	Sn(-) %	Sp(-) %	FPR %	FPR TM# %
TPpred*	94	0.73	73	81	97	96	2.6	15.5
PCLR°	86	0.6	93	48	84	99	0	12.8

Gli indici di performance sono descritti alla sezione "Calcolo delle performance" a pagina 11. La valutazione riportata in tabella comprende 297 esempi positivi e 8010 esempi negativi. *, ° e # come descritti nella Tabella 4.

Predizione del sito di taglio

Oltre che la per la predizione della presenza del Targeting Peptide, il predittore TPpred è stato sviluppato per identificare anche la posizione del sito di taglio lungo la sequenza analizzata. Questa informazione è molto importante dal momento che permette di conoscere qual'è la sequenza della proteina matura dopo l'avvenuta modificazione post-traduzionale. Le prestazioni di TPpred nella predizione del corretto sito di taglio sono state valutate in maniera comparativa con quelle ottenute con i predittori TargetP, PredSL and MitoProt (quest'ultimo solo per le sequenze mitocondriali) e non con i predittori Predotar, iPSORT e PCLR in quanto non è prevista la predizione del sito di taglio per questi ultimi. Sequenze mitocondriali e plastidiche sono state valutate separatamente data la rilevante differenza in termini di lunghezze medie (22 e 59 residui rispettivamente). Per ciascuna sequenza testata abbiamo calcolato l'errore come la differenza tra la posizione reale del sito di taglio e quella

predetta. In seguito sono stati valutati due diversi indici: l'errore medio (ME) e il numero di sequenze predette per le quali l'errore è più basso della deviazione standard (σ) della distribuzione delle lunghezze dei Targeting Peptide (E<σScore), ovvero E<σ=16 per Targeting Peptide mitocondriali e E<σ=22 per Targeting Peptide cloroplastici. TPpred risulta il miglior predittore di sito di taglio dall'osservazione di entrambi gli indici ME e E<σScore e questo significa che il predittore è in grado di predire correttamente la posizione del sito di taglio anche se la distribuzione delle lunghezze è molto variabile (Tabella 6).

Tabella 6. Confronto sulla predizione del sito di taglio del Targeting Peptide.

	Mitocondri		**Cloroplasti**	
Metodo	**ME (res)**	**E<σscore (%)**	**ME (res)2**	**E<σ score (%)**
TPpred *	7	89	15	74
TargetP °	12	71	16	71
PredSL °	12	75	17	73
MitoProt °	13	75	-	-

ME: errore medio nella predizione del sito di taglio. E<σ score: percentuale di predizioni con errore più basso della deviazione standard della distribuzione delle lunghezze de Targeting Peptide. * e ° come in Tabella 4.

Predizione di interi proteomi

TPpred è stato impiegato per la predizione del targeting peptide dell'intero proteoma di tre differenti specie eucariotiche recuperate dalla banca dati EnsEMBL (Human, Arabidopsis e yeast). I risultati delle predizioni sono riportati nella Tabella 7. Abbiamo stimato che rispettivamente il 4,0%. 9,0% e il 6,1% delle sequenze proteiche testate per uomo, pianta e lievito risultano dotate di Targeting Peptide. Bisogna notare che tale stima è più bassa rispetto quella riportata precedentemente nella letteratura (10-25%, Emanuelsson et al., 2000). Tale dato è la risultante dell'effetto combinato di due fattori:

- Il basso tasso di falsi positivi dato da TPpred (FPR=3%) che limita il numero di predizioni erronee all'interno del set di esempi negativi;
- La sensitività di TPpred (Sp+=75%) che implica il fatto che il 25% delle sequenze dotate di Targeting Peptide non vengano rilevate

Tuttavia, dal momento che il set negativo è enormemente più grande rispetto all'insieme di sequenze dotate di Targeting Peptide, quando si debbono predire interi proteomi è preferibile utilizzare un metodo con performance migliori in termini di FPR.

Tabella 7. Predizione del Targeting Peptide in interi proteomi

	Homo sapiens	**Arabidopsis thaliana**	**Saccharomyces cerevisiae**
Intero Proteoma # Proteine	93,588	35,386	6,692
Targeting Peptide Predetto# Proteine	3,744	3,194	407

Risultati SPpred

Caratteristiche dei Signal Peptide

Lunghezza: La lunghezza dei Signal Peptide di eucarioti, batteri Gram+ e Gram- inclusi nel nostro dataset varia da 15 a 50 residui, con medie rispettive di 23.3, 31.3 e 24.5 (Figura 5 A-B-C). I Signal Peptide di Gram+ risultano essere mediamente più lunghi rispetto quelli di sequenze appartenenti agli altri due gruppi per giunta con una deviazione standard molto più ridotta (5.6, 6.0 e 5.9 rispettivamente per eucarioti, batteri Gram+ e Gram-). A partire da questa prima osservazione in merito alla distribuzione delle lunghezze, si è deciso di separare i due dataset batterici in Gram+ e Gram-, e di trattare separatamente l'addestramento del GRHCRF.

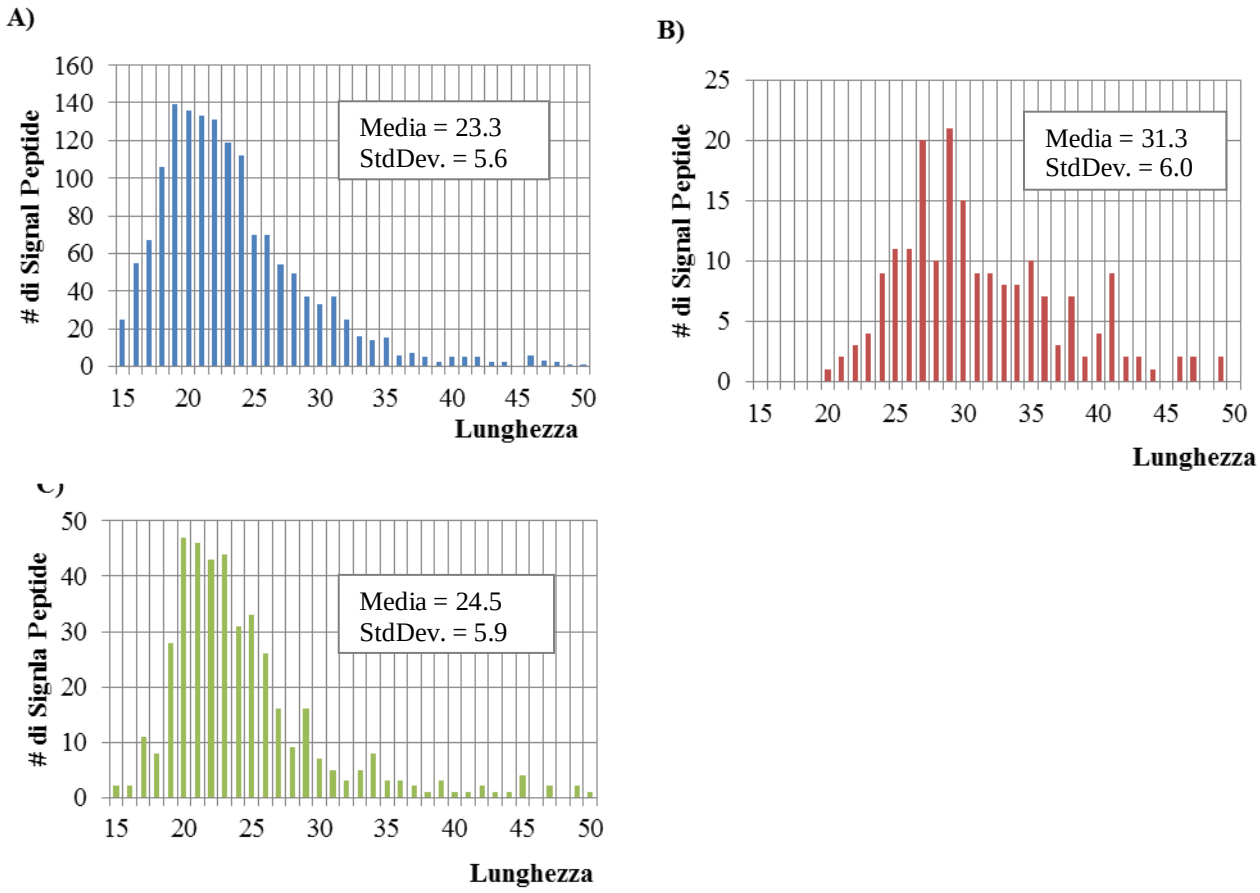


Figura 5. Distribuzione della lunghezza del Signal Peptide del dataset di esempi positivi di Eucarioti (A), Gram+ (B) e Gram- (C).

Composizione dei residui: La composizione dei residui dei Signal peptide è mostrata nella Figura 6 nella quale vengono confrontate le frequenze amminoacidiche dei Signal Peptide con quelle delle intere sequenze incluse nei tre dataset. È possibile notare che non si riscontrano differenze notevoli se si confrontano le composizioni relative alle sequenze senza Signal Peptide (in giallo) con quelle calcolate sulle intere sequenze di proteine dotate di Signal Peptide (in rosso), mentre risulta decisamente caratteristica la composizione dei Signal Peptide (in azzurro). In particolare si nota che i Signal Peptide sono ricchi di amminoacidi idrofobici come Alanina (A), Leucina (L), Metionina (M) e Fenilalanina (F), mentre risultano nettamente carenti di residui carichi negativamente (D, E) e di Tirosina (Y), inoltre, anche se misura minore nei Signal Peptide scarseggiano i residui carichi positivamente (R, K), Istidina (H), Glutammina (Q) e Asparagina (N).

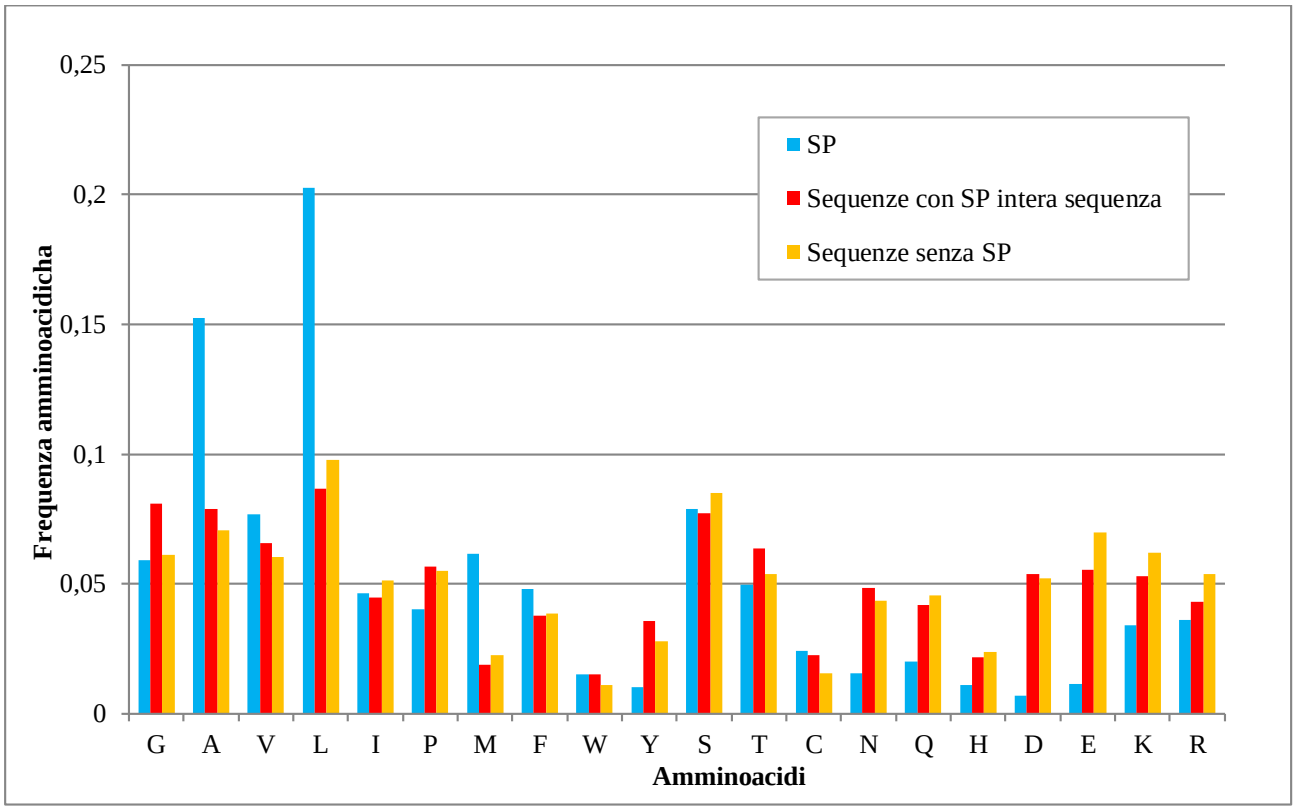


Figura6: Composizione amminoacidica dei Signal Peptide (azzurro), delle intere sequenze dotate di Signal Peptide (rosso) e delle sequenze sprovviste di Signal Peptide (giallo).

Sito di taglio: nella regione del Signal Peptide adiacente al sito di taglio risiede il segnale di riconoscimento per gli enzimi peptidasici, ci aspettiamo, pertanto, che sia riconoscibile nell'intorno di tale sito un pattern specifico, o perlomeno, una regione altamente conservata. Le proteine dei tre dataset positivi sono state allineate separatamente considerando i 20 residui a valle e i 10 a monte del sito di taglio e per ciascun allineamento è stata effettuata la rappresentazione mediante "sequence

logo" (Figura 7 A-B-C). Dalla rappresentazione è possibile evidenziare quali sono le regioni maggiormente conservate intorno al sito di taglio (che risiede tra le posizioni -1 e 0): per tutti tre i gruppi è possibile notare che il residuo subito prima del sito di taglio (posizione -1) è molto conservato così come il residuo in posizione -3, in entrambi i casi l'amminoacido più frequente risulta essere una Alanina (A). Le posizioni a valle del sito di taglio, invece, appaiono molto più eterogeneamente assortite per le sequenze eucariotiche, mentre per i batteri è interessante notare una discreta frequenza di amminoacidi a carica negativa (D e E) nelle posizioni 0 e 1. Osservando le Figure 7A, 7B e 7C è possibile anche individuare una regione fortemente idrofobica (con predominanza di L, A e V) circa 10-15 residui a monte del sito di taglio. Inoltre, per il gruppo Gram- e, in misura minore, per gli eucarioti si nota una regione ricca di Lisina (K) intorno alla posizione -18.

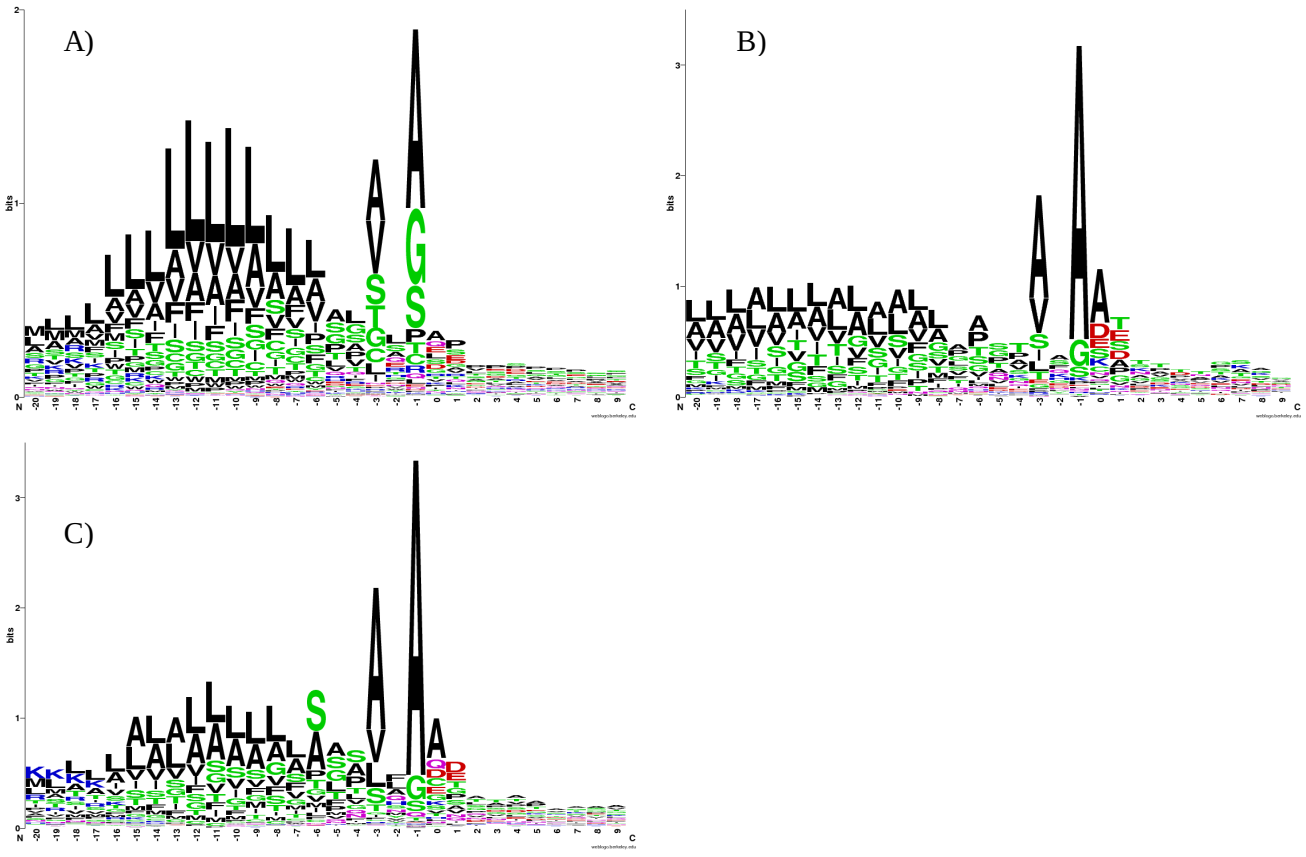


Figura 7. Sequence logo delle posizioni intorno al sito di taglio dei Signal Peptide. Il sequence logo è stato calcolato per mezzo dell'applicazione online WebLogo (http://weblogo.berkeley.edu/). La posizione "0" indica il primo residuo della sequenza matura. L'altezza delle lettere è proporzionale alla frequenza dell'amminoacido nella posizione corrispondente. I codice colore è organizzato per distinguere i residui idrofobici (nero), polari (verdi), carichi positivamente (blu) e carichi negativamente (rosso). I tre riquadri A, B e C si riferiscono rispettivamente alle sequenze di Eucarioti, Gram+ e Gram-.

Modello GRHCRF per SPpred

Le caratteristiche delle sequenze peptidiche relative ai Signal Peptide, che sono state esaminate nei paragrafi precedenti, sono state tenute in considerazione nella progettazione dei modelli predittivi che vengono rappresentati schematicamente nella Figura 8. Tre modelli sensibilmente differenti sono stati impiegati per i tre gruppi di sequenze: il modello base (senza le transizioni rappresentate con le frecce rosse) è il modello predittivo di batteri Gram+, il modello per batteri Gram- aggiunge le transizione rappresentate dalle frecce rosse a linea continua, infine il modello per sequenze eucariotiche aggiunge le transizioni rappresentate da tutte le frecce rosse (a linea continua e tratteggiata). Nonostante queste piccole differenze, resesi necessarie per modellizzare le diverse lunghezze e composizioni dei Signal Peptide, è possibile descrivere uno schema comune per tutti tre gli automi. Dallo stato di inizio (Begin) sono possibili due transizioni verso due sottomodelli che descrivono le sequenze dotate di Signal Peptide (sottomodello signal, in blu) e le sequenze sprovviste di Signal Peptide (sottomodello non-signal, in arancio). I due sottomodelli sono identici nell'architettura, ovvero presentano esattamente lo stesso numero di stati e le medesime transizioni. Il sottomodello signal è composto da quattro moduli:

- Una regione N-terminale, composta da 5 stati, atto a catturare l'ipervariabilità dei residui N-terminali e la variabilità della lunghezza di tale regione;
- Un dominio centrale, composto da 10 stati, volto a modellizzare la regione idrofobica resa nota dall'analisi composizionale effettuata nei paragrafi precedenti;
- Un dominio successivo, composto da 5 stati, per convogliare la porzione di sequenza ad alta variabilità che è presente prima della regione del sito di taglio;
- La regione del sito di taglio con 3 stati monte e tre a valle obbligatori.

Durante la fase di ottimizzazione si è notato che il modello così descritto, pur essendo dell'architettura ottimale per predire la presenza del Signal Peptide, risultava insoddisfacente nella predizione del corretto sito di taglio. È stato pensato quindi di introdurre una seconda fase predittiva per tutte le sequenze predette come aventi Signal Peptide per migliorare la collocazione del sito di taglio. A questo scopo si utilizza, per ciascuno dei gruppi di sequenze, un ulteriore modello, addestrato unicamente con sequenze dei dataset positivi, mostrato nella Figura 8, in cui gli stati blu portano la label "Y" (signal) e gli stati arancio la label "N" (non-signal).

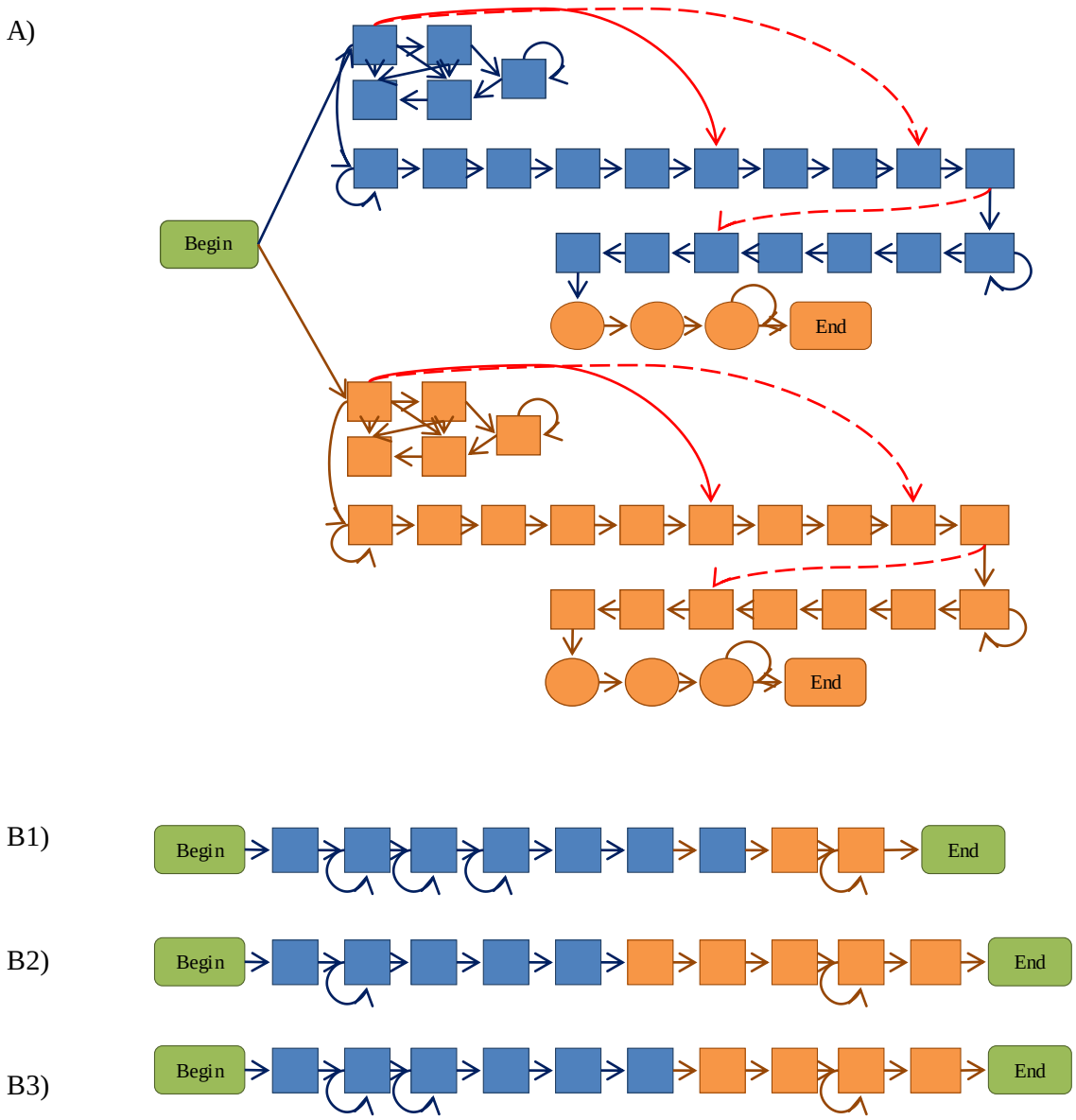


Figura 8. Modello predittivo di Signal Peptide. **G**li stati blu rappresentano la label "Y" (Signal peptide) mentre quelli in arancio rappresentano gli stati con label "N" (non-signal). A) Per Gram+, Gram- ed Eucarioti vengono utilizzati tre modelli differenti: il modello di base è relativo ai Gram+, i link rappresentati dalle frecce rosse continue sono aggiunti nel modello relativo ai Gram-, mentre due ulteriori link (rappresentati dalle frecce rosse tratteggiate) sono introdotti nel modello relativo ad Eucarioti. B) Tre diversi modelli per l'identificazione del sito di taglio: B1) Gram+, B2) Gram+, B) Eucarioti.

Predizione con differenti input

Il modelli descritti nel paragrafo precedente sono stati addestrati adottando la strategia descritta nel paragrafo "Addestramento, validazione e test" di pagina 11 per mezzo di una procedura 5-fold cross validation: per ciascun gruppo di sequenze (Gram+, Gram- ed Eucarioti) il dataset è stato suddiviso in 5 sottoinsiemi dei quali 3 sono stati impiegati per il training, 1 per scegliere automaticamente i parametri del modello (set di validazione) e 1 per calcolare gli indici di performance (set di test). Dal momento che i dataset sono stati globalmente ridotti per omologia (si veda paragrafo "Dataset di Signal Peptide" di pagina 9) tale procedura assicura risultati attendibili ed evita il problema dell'overfitting.

La Tabella 8 mostra le prestazioni del predittore SPpred sul set completo di sequenze comprese nei tre dataset a seconda della tipologia di input adottata. Utilizzando come input solamente la singola sequenza si ottiene un'accuratezza globale (Acc) del 98% e un Matthews Correlation Coefficient (MCC) di 0.89. Aggiungendo a mano a mano le differenti caratteristiche descritte nel paragrafo "Input" di pagina 11 le prestazioni del predittore aumentano sensibilmente fino a raggiungere i massimi valori quando l'input è composto da sequenza, idrofobicità e carica. Aggiungendo all'input la feature relativa al momento idrofobico non si evincono miglioramenti nelle prestazioni. Nonostante ciò si è deciso di utilizzare l'input completo per uniformarsi a quanto già descritto per il predittore TPpred, nell'ottica futura di potere implementare i due predittori in un unico tool bioinformatico. Le performance finali di SPpred ammontano al 98% di Acc e a 0.90 di MCC.

Tabella 8. Performance di GRHCRF con diverse codifiche di input per la predizione del Signal Peptide.

Input	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(-) (%)	Sp(-) (%)	FPR (%)
seq	98	0,89	90	90	99	99	1,2
seq+kd	98	0,89	91	90	99	99	1,1
seq+kd+ch	98	0,90	91	91	99	99	1,1
seq+kd+ch+hm (SPpred)	98	0,90	91	91	99	99	1,1

seq: sequenza; kd: livello di idrofobicità calcolato con la scala di Kyte-Doolittle; ch: carica ; hm: momento idrofobico. La valutazione riportata in tabella include 2106 esempi positivi e 18377 esempi negativi che comprendono sequenze di Eucarioti, Gram+ e Gram-. Gli indici di performance sono calcolati con una procedura di 5-fold cross-validation.

Confronto con i predittori allo stato dell'arte

La Tabella 9 riporta le prestazione di SPpred a confronto con i metodi attualmente disponibili per la predizione di Signal Peptide. È possibile constatare che l'ultima versione di SignalP presenta un MCC di poco superiore a quello ottenuto con SPpred (0.91 contro 0.90) con un indice di accuratezza globale identico. Tuttavia, in termini di tasso di falsi positivi (FPR), SPpred risulta il predittore migliore.
Inoltre va considerato che, mentre le performance di SPpred sono valutate in cross-validation, tutti gli altri predittori sono testati con un set di sequenze non indipendente dai relativi set di training. SPpred risulta ben bilanciato tra specificità e sensitività di entrambe le classi predittive a differenza di tutti gli altri predittori che presentano una sensitività alta a discapito della specificità, ragione per cui SPpred ha il più basso FPR. Un altro indice di valutazione estremamente importante è il FPR nel sottoinsieme di sequenze senza Signal Peptide che contengono un'elica transmembrana nella regione N-terminale.

Tabella 9. Risultati ottenuti con TPpred a confronto con i predittori allo stato dell'arte non specifici per mitocondri o cloroplasti

METODO	Acc %	MCC	Sn(+) %	Sp(+) %	Sn(-) %	Sp(-) %	FPR %	FPR TM %#
SPpred*	98	0,90	91	91	99	99	1,1	4,3
SignalP4.1°	98	0,91	94	89	99	99	1,3	4,6
SignalP4.0°	98	0,90	94	88	98	99	1,5	4,7
Spoctopus°	95	0,78	89	73	96	99	3,8	12,9
Phobius°	94	0,75	96	63	93	100	6,5	17,4
POLYPhobius°	94	0,74	88	67	95	99	5,1	12,5
Philius°	92	0,70	96	57	92	99	8,2	12,5
PREDisi°	92	0,68	93	56	92	99	8,4	40,2
SPEPLip°	86	0,51	82	41	86	98	13,6	38,2
Signal-3L°	81	0,50	95	34	79	99	21,1	75,0
Signal-CF°	80	0,50	97	34	78	100	21,8	77,7
Signal Blast°	77	0,48	99	31	75	100	25,0	53,3

Gli indici di performance sono descritti nel paragrafo "Calcolo delle performance" a pagina 11. La valutazione riportata in tabella include 2106 esempi positivi e 18377 esempi negativi che comprendono sequenze di Eucarioti, Gram+ e Gram-. * Le sequenze incluse nel set di training sono predette in cross-validation, le sequenze del dataset negativo non incluse nel dataset di training sono predette selezionando in modo casuale uno dei 5 GRHCRF addestrati. ° Il dataset di confronto si sovrappone con il dataset di training. # Il tasso di falsi positivi (FPR) è calcolato sul set di esempi negativi che sono dotati di una alpha-elica transmembrana nei 60 residui N-terminali (Tabella 2).

Data l'elevata idrofobicità nella composizione amminoacidica dei Signal Peptide e delle eliche transmembrana è possibile che un predittore confonda le due tipologie di strutture, ovvero che sequenze contenenti un alpha-elica vengano predette come aventi Signal Peptide. Dalla tabella si può

notare che con SPpred si ottiene il più basso FPR per il sottogruppo di sequenze transmembrana, questo rende il nostro predittore il più a adatto per effettuare questo tipo di discriminazione.

Predizione del sito di taglio

Oltre che la per la predizione della presenza del Signal Peptide, il predittore SPpred è stato ideato per identificare anche la posizione del sito di taglio lungo la sequenza analizzata. Avere indicazioni sulla posizione di tale sito è di notevole importanza se si desidera determinare la sequenza matura delle proteine testate. Le performance di SPpred nella predizione del corretto sito di taglio sono state confrontate con quelle ottenute con i predittori allo stato dell'arte già descritti nei paragrafi precedenti. I tre gruppi di sequenze analizzate (Eucarioti, Gram+ e Gram-) sono stati valutati in maniera congiunta in quanto non si rilevano grandi differenze in termini di lunghezze medie e deviazioni standard (si veda il paragrafo "Caratteristiche dei Signal Peptide" di pagina 22).
La valutazione delle prestazione viene fatta calcolando Accuratezza Globale (ACC) e Coefficiente di Correlazione (MCC) considerando come veri positivi (TP) solamente le sequenze in cui l'errore nella predizione dei sito di taglio è più basso della deviazione standard (errore definito come il valore assoluto della differenza tra la posizione reale del sito di taglio e quella predetta). SPpred, testato in cross-validation, presenta una ACC di 98% e un MCC di 0.88 che risultano essere le performance migliori tra tutti i predittori testati fatta eccezione per SignalP4.1 che presenta un MCC superiore di un solo punto percentuale (0.89). Tali risultati sono mostrati nella Tabella 10.

Tabella 10. Confronto sulla predizione del sito di taglio del Targeting Peptide.

METODO	Acc %	MCC
SPpred	98%	0,88
SignalP4.1	98%	0,89
SignalP	98%	0,88
Spoctopus	95%	0,75
POLYPhobius	94%	0,71
Phobius	93%	0,71
Philius	92%	0,67
PREDisi	91%	0,66
SPEPLip	84%	0,40
Signal-3L	80%	0,47
Signal-CF	79%	0,44
Signal Blast	77%	0,46

Predizione di interi proteomi

Analogamente a quanto già descritto per il predittore TPpred, anche SPpred è stato impiegato per la predizione del Signal Peptide dell'intero proteoma di tre differenti specie, ciascuna rappresentate uno dei sottogruppi con cui è stato addestrato il tool: Homo sapiens per gli Eucarioti, Escherichia coli per i Gram- e Clostridium sticklandii per i Gram+. Tutti i proteomi sono stati recuperati dalla banca dati EnsEMBL e la numerosità di ciascun proteoma è riportata nella Tabella 11 insieme al numero di sequenze predette come aventi Signal Peptide. Da questi valori abbiamo stimato che rispettivamente il 12,0%, 11,3% e 5,5% delle sequenze testate per H.sapiens, E.coli e C.sticklandii risultano dotate si Signal Peptide. Anche in questo caso, come per la predizione di interi proteomi di Targeting peptide, bisogna notare che la stima è più bassa di quella riportata precedentemente nella letteratura e le ragioni sono riconducibili al fatto che SPpred è in grado di limitare il numero di falsi positivi dato il basso FPR (1,1%) e la capacità di discriminare le alpha-eliche transmembrana dai Signal Peptide (FPR TM =4,3%).

Tabella 11. Predizione del Signal Peptide in interi proteomi

	Homo sapiens	**Escherichia coli**	**Clostridium sticklandii**
Intero Proteoma # Proteine	93,588	4,653	2,571
Signal Peptide Predetto #Proteine	11,265	526	142

CONCLUSIONI

Durante lo svolgimento del Dottorato di Ricerca sono stati implementati due predittori per Targeting Peptide e Signal Peptide, rispettivamente denominati TPpred e SPpred.

Entrambi i tool sono basati sulla metodologia di machine-learning GRHCRF, introdotta recentemente per la risoluzione di problemi di labelling di sequenze biologiche. I predittori sono stati addestrati con Targeting e Signal Peptide determinati sperimentalmente e introducendo, ove necessario, procedure di selezione al fine di bilanciare i dataset di esempi positivi e negativi con lo scopo finale di aumentare la capacità discriminativa dei metodi.

TPpred risulta essere il migliore predittore tra quelli allo stato dell'arte sia per la predizione della presenza di Targeting Peptide mitocondriali o cloroplastici sia per la determinazione del corretto sito di taglio. Anche per SPpred è possibile effettuare un buon confronto con gli altri predittori, tuttavia le prestazioni del metodo SignalP4.1 risultano uguali o superiori rispetto a quelle date dal nostro predittore, motivo che ci incoraggia a proseguire il lavoro qui presentato.

La caratteristica peculiare di TPpred e SPpred è che entrambi risultano essere molto più specifici rispetto a tutti gli altri predittori, dando un basso tasso di falsi positivi. Questa caratteristica rende i due oggetti adatti alla predizione di Targeting Peptide e di Signal Peptide in interi proteomi.

ANNOTAZIONE DI GENOMI: **Analisi di dati Next Generation Sequencing nella ricerca sul cancro**

BACKGROUND

Il secondo ambito di studi svolto durante il Dottorato di Ricerca è consistito in una serie di lavori di carattere più applicativo, condotti su dati genomici ricavati da pazienti reali, che dimostrano l'ampia necessità di applicazione di strumenti bioinformatici di analisi e predizione nell'interpretazione dei dati che le tecnologie attuali mettono in grado di ottenere. I casi trattati sono stati molteplici e in qualche misura eterogenei e nel presente elaborato si esporrà solamente quello a mio parere più significativo, fermo restando che in appendice sono disponibili tutte le pubblicazioni in cui le ricerche sono descritte.
In particolare tutti i progetti hanno riguardato l'analisi di dati prodotti mediante tecnologia Next generation Sequencing (NGS), per cui si è messa a punto una metodologia di analisi bioinformatica per dati di RNA-seq e Whole ExomeSequencing (WES).
Le tecnologie NGS rappresentano un potente mezzo nell'ambito della ricerca genomica in quanto per mezzo di esse è possibile sequenziare l'intero genoma, esoma o trascrittoma un individuo in tempi e produttività competitive rispetto i metodi tradizionali.
L'avvento di tali tecnologie pone però il mondo della ricerca di fronte al problema dell'analisi dei dati, aspetto che può essere approcciato unicamente con competenze informatiche e consistenti capacità di calcolo.
Presso il Centro Interdipartimentale di Ricerche sul Cancro "Giorgio Prodi" (CIRC) dell'Università di Bologna si svolgono, da quattro anni a questa parte, ricerche che vengono condotte sulla piattaforma NGS Illumina HiScanSQ. In particolare sono stati condotti esperimenti di RNA-seq e WES su numerosi campioni oncologici e non, relativi ad altrettanto numerosi progetti di ricerca. In generale ciascuno di tali progetti ha come scopo finale quello di rilevare lesioni genetiche presenti nel campione correlabili alla malattia del paziente in modo da poter individuare nuovi target diagnostici e terapeutici.
In questo contesto l'analisi computazionale rappresenta un passaggio chiave.
Per la messa a punto della pipeline di analisi sono state impiegate le risorse bioinformatiche allo stato dell'arte come algoritmi di allineamento, software per la detection di varianti come "Single Nucleotide Variants" (SNV) e riarrangiamenti cromosomici, strumenti per lo studio dell'espressione genica, metodi di annotazione genomica. Inoltre sono stati implementati numerosi tool ad hoc sfruttando le proprie conoscenze informatiche nei linguaggi di programmazione Python, awk e Perl.
Il metodo sviluppato è stato applicato a numerosi campioni oncologici di diversa tipologia. In particolare si è dato il proprio contributo in studi inerenti: Leucemie Mieloidi Acute pediatriche (LAM-infant), Tumori Stromali Gastro-Intestinali (GIST), Adenocarcinomi Pancreatici (PDAC), Mioepiteliomi, Sarcomi a cellule chiare del rene (CCSK), Liposarcomi, Sarcomi di Ewing, Condrosarcomi Mixoidi Extraschelertrici (EMCS), Dermatofibrosarcoma Protuberans (DFSP), Acalasia.

PIPELINE DI ANALISI NGS

Come già introdotto, nella presente trattazione si andrà ad introdurre il metodo di analisi di dati NGS adottata nelle applicazioni di ricerca sul cancro illustrando brevemente tutti gli step in cui vengono adottati applicativi bioinformatici noti o metodologie sviluppate in casa. La Figura 1 mostra, in maniera schematica, le operazioni che sono state implementate nella pipeline di analisi.

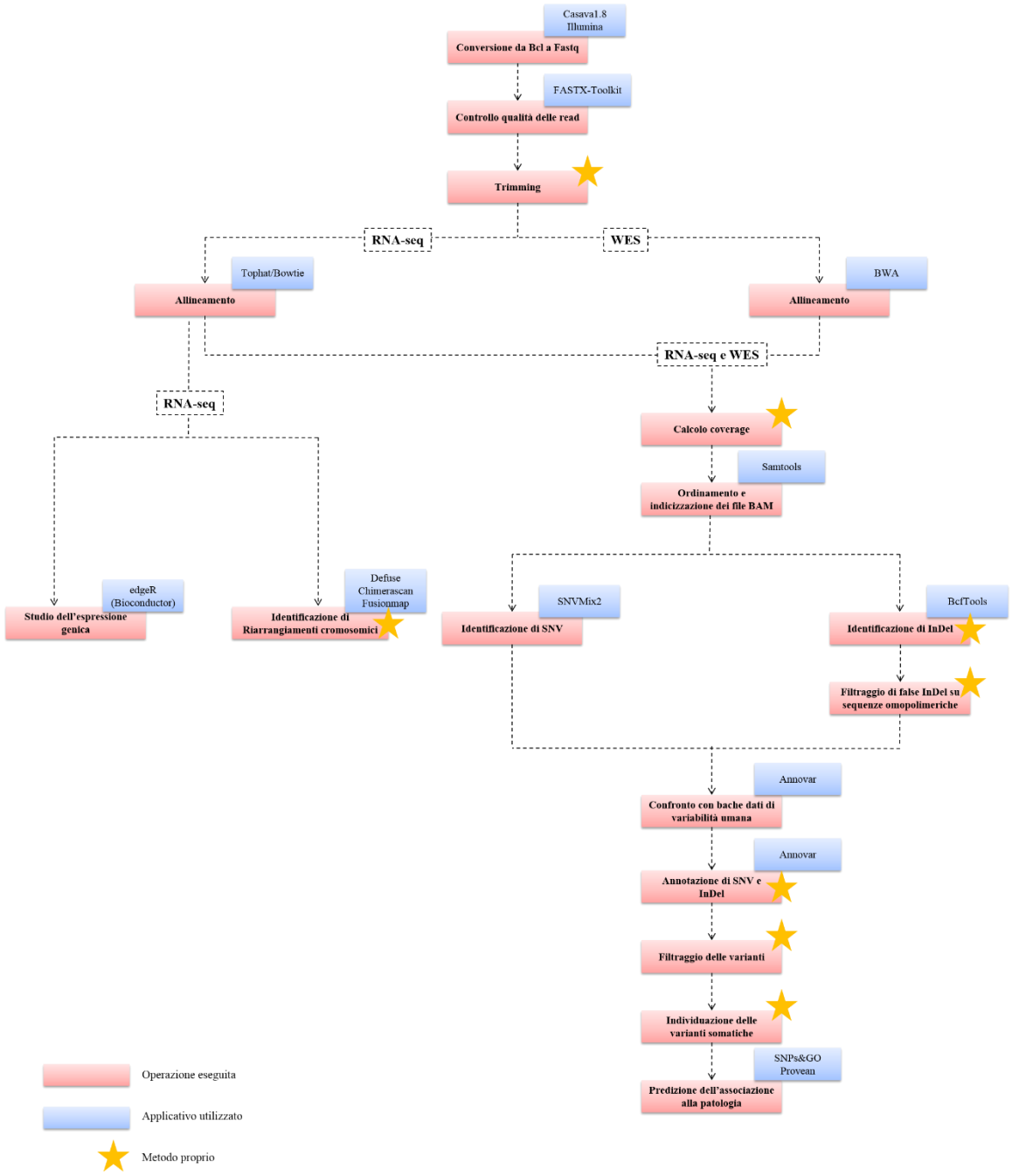


Figura 1. Rappresentazione schematica della pipeline di analisi di dato NGS

- ***Conversione da Bcl a Fastq.*** Il primo step dell'analisi bioinformatica consiste nella conversione del dato grezzo, fornito dalla piattaforma di sequenziamento (nel nostro caso dal formato non human-readable denominato Bcl, specifico di Illumina), al dato di sequenza nel formato universalmente riconosciuto Fastq (Figura 2). Questa operazione viene condotta con l'applicativo sviluppato da Illumina Casava1.8[36]. Spesso, alla fase di conversione, si associa lo step di demultiplexing. Tale step è necessario quando, data l'elevata produttività della piattaforma utilizzata, si decide di sequenziare all'interno dello stesso ambiente di reazione (lane) più di un campione contemporaneamente. Le sequenze vengono ricondotte al corretto campione grazie alle diverse sequenze nucleotidiche (index o barcode) che vengono aggiunte in fase di preparazione delle librerie.

```
@Unique Identifier Illumina
GGTGAAGCAAGTCACGTGATCACAGGACAGCGGGCCCTTCCCTCTTAGGTAGCTCCAGATGCTGCAGCAGCAGGA
+
]U]YS]]]]M[[QVL[OONN\TNNY]QZQZZXX]]_\___BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Figura 2. Formato fastq. Ciascuna sequenza viene identificata quattro righe: la prima contiene l'identificativo unico ed inizia con il carattere "@", la seconda contiene la sequenza nucleotidica, la terza contiene il carattere "+" e l'ultima contiene una stringa che codifica (in codice ACSII come Phred quality) la qualità di ciascuna base sequenziata.

- ***Controllo della qualità delle read.*** Durante il sequenziamento ad ogni base letta viene associato un valore numerico che indica la qualità della lettura. L'andamento della qualità è un indice molto importante per valutare la buona riuscita della run di sequenziamento. Di norma, in prossimità degli ultimi nucleotidi sequenziati, il valore di qualità ha un decadimento che va valutato per decidere quale porzione di sequenza è opportuno considerare negli step succesivi. L'applicativo che viene utilizzato a tale proposito è la funzione *fastx_quality_stats* appartenente al pacchetto FASTX-Toolkit[37].

- ***Trimming.*** Se mediante lo step di controllo qualità viene decretato che gli ultimi nucleotidi letti hanno valori mediamente bassi si procede con la procedura di trimming, ovvero la rimozione dell'ultima o delle ultime basi sequenziale. In questo modo la sequenza che viene impiegata per gli step successivi risulterà più corta ma con una elevata qualità in ogni sua parte. Questo passaggio viene eseguito per mezzo di un semplice script elaborato nel linguaggio di programmazione Python.

- ***Allineamento.*** Questo step rappresenta la fase più onerosa dal punto di vista computazionale. Come è possibile notare nella Figura 1 il software impiegato a tale proposito è diverso se stiamo trattando dati di RNA-seq o di WES. Nel primo caso infatti è necessario utilizzare un applicativo che preveda l'utilizzo di annotazioni geniche (file GTF – Gene Transfer Format) e che permetta di effettuare allineamenti sulle giunzioni esoniche. Viene impiegato a tale proposito la pipeline Tophat/Bowtie[38]. Nel caso di dati WES, non essendo necessarie specificità accennate per l'RNA-seq, è stato scelto il software BWA[39]. In entrambi i casi gli allineamento sono condotti contro il Genoma Umano di riferimento HG19 (disponibile al link http://genome.ucsc.edu/).

- ***Calcolo della profondità di lettura media.*** Nella Figura 1 è lo step che viene indicato come "Calcolo della Coverage". "Profondità di lettura" e "Coverage" sono infatti termini che spesso vengono utilizzati per designare il numero di sequenze che ricoprono una certa regione genomica. Il calcolo viene eseguito con uno script programmato nei linguaggi Python ed Awk.

- ***Ordinamento e Indicizzazione.*** I tool impiegati per l'allineamento restituiscono come output un file BAM, ovvero un particolare formato atto a contenere le informazioni su ciascuna sequenza mappata contro un genoma di riferimento. Tale file può subire diverse manipolazioni tra cui l'ordinamento delle sequenze contenutevi in base alla coordinata genomica e l'indicizzazione. A questo proposito viene impiegato l'applicativo *samtools* con le funzioni *sort* e *index*[40]. Oltre che ordinamento ed indicizzazione, l'applicativo *samtools* viene utilizzato per trasformare l'allineamento dal formato BAM al formato PILEUP, utile nei passaggi successivi di identificazione di SNV e Indel.

- ***Studio dell'espressione genica***. Negli studi condotti mediante metodologia RNA-seq l'obbiettivo finale è, talvolta, la valutazione comparativa dell'espressione genica tra due o più condizioni diverse (come per esempio gruppi di pazienti o pazienti trattati con diversi farmaci). L'espressione di un gene viene valutata in base alla profondità di lettura che gli viene attribuita, ovvero al numero di frammenti sequenziati che sono stati allineati al locus genomico di interesse. Viene quindi effettuato un primo step di conteggio delle read mappanti su ciascun gene per mezzo della funzione Python htseq-count. In seguito tale dato viene analizzato con il pacchetto *edgeR* di R-Bioconductor[41] che implementa il modello Binomiale Negativo.

- ***Identificazione di riarrangiamenti cromosomici***. Per mezzo di RNA-seq è possibile effettuare l'identificazione di riarrangiamenti cromosomici (traslocazioni, inversioni, delezioni o trasposizioni) che danno origine a trascritti di fusione. Nella pipeline implementata si utilizzano tre diversi applicativi bioinformatici che, a partire dai file Fastq, effettuano l'identificazione di

tali eventi: deFuse[42], Chimerascan[43] e FusionMap[44]. L'output finale che viene considerato è costituito dall'unione delle intersezioni a coppie dei trascritti di fusione individuati da ciascun software:

(Defuse ∩ Chimerascan) | (Defuse ∩ FusionMap) | (FusionMap ∩ Chimerascan)

- ***Identificazione di varianti a singolo nucleotide.*** Lo step di identificazione di SNV viene effettuato a partire dall'osservazione nell'allineamento nelle coordinate genomiche in cui viene letta una base diversa da quella presente nel genoma di riferimento (Figura 3). La variante osservata può essere omozigote o eterozigote (nella Figura 3 viene mostrata una variante omozigote). Tuttavia nei campioni tumorali è possibile che ci sia una ploidia alterata o una contaminazione da tessuto non oncologico che vanno ad impattare la significatività statistica delle varianti detectate. Per questa ragione viene impiegato un applicativo specifico per l'identificazione di varianti a singolo nucleotide in studi oncologici: SNVMix2[45].

Figura 3. Esempio di allineamento e identificazione di una variante a singolo nucleotide A>T

- ***Identificazione di inserzioni e delezioni***. Inserzioni e delezioni vengono individuate per mezzo di BcfTools[40] il cui output viene elaborato per mezzo di uno script implementato in Python. Un'attenzione particolare si è riservata allo sviluppo di un metodo di riconoscimento e filtraggio delle inserzioni in regioni omopolimeriche (di tipo AAA, TTT, GGG o CCC). Da un'indagine

sui dati è emerso che la piattaforma Illumina commette questo tipo di errore principalmente in omoesameri di Adenina e Timina. È stato quindi individuato di un filtro, basato sulla coverage dell'inserzione e sulla lunghezza dell'omopolimero, atto ad eliminare i falsi positivi lasciando tra le InDel candidate solamente quelle che con alta probabilità sono mutazioni effettive del campione.

- ***Confronto con banche dati di variabilità umana.*** Negli studi condotti in campo onco-ematologico, si ricercano varianti nuove o rare che possano essere significativamente associate alla patologia. È quindi necessario uno step di confronto con banche dati di variabilità umana che permetta di filtrare tutte le varianti che risultano essere polimorfismi (frequenza nella popolazione generale > 1%). Il confronto viene effettuato con le banche dati 1000genomes (http://www.1000genomes.org/) e dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) nelle release più aggiornate, utilizzanzo il software Annovar[46].

- ***Annotazione delle varianti.*** L'applicativo Annovar viene utilizzato anche per la fase di annotazione. Ciascuna variante trovata (SNV o Indel) viene mappata all'interno del gene e della rispettiva sequenza proteica per individuare tipo di sostituzione amminoacidica (nel caso delle SNV) o l'effetto di frameshift (nel caso delle Indel).

- ***Filtraggio delle varianti***. In seguito alle fasi sopra riportate, le varianti trovate vengono sottoposte ad uno step di filtraggio atto ad eliminare il maggior numero possibile di falsi positivi. Sono stati individuati due criteri di filtraggio, entrambi basati sul concetto di coverage:1) Coverage totale, intesa come numero di read in cui viene letta mutazione (base alterata o Indel) sommata al numero di read che supportano la condizione wild-type; 2) Ratio, ovvero il rapporto tra il numero di read che supportano la mutazione e la Coverge totale. È opportuno che la soglia di Coverage totale scelta sia proporzionata alla Coverage media del campione, in generale non sotto un valore di 10. La soglia di Ratio viene fissata arbitrariamente a 0.3 o talvolta più bassa quando si presume una consistente presenza di tessuto normale nel campione analizzato. È stato riscontrato che, utilizzando le soglie indicate, siamo in grado di evidenziare varianti in geni o in alleli poco espressi e in campioni con cellularità tumorale variabile.

- ***Individuazione delle varianti somatiche***. Applicando la pipeline descritta al campione tumorale e alla rispettiva controparte normale (di norma analizzata per mezzo di WES) è possibile effettuare un confronto che permette di distinguere quali sono le varianti somatiche (ovvero proprie del tumore) e quali sono germinali (ovvero riconducibili alla variabilità individuale del paziente analizzato).

- ***Predizione dell'associazione alla patologia.*** Giunti a questo punto della pipeline di analisi, nonostante i criteri di filtraggio e ottimizzazione adottati, possono essere ancora molte le varianti che emergono. Per questo motivo nasce l'esigenza di applicare un metodo che consenta la prioritizzazione degli eventi individuati. Nel nostro workflow vengono introdotti tre step predittivi con tre diversi tool: SNPs&GO[47], PROVEAN[48] e I-Mutant[49]. Così come i predittori TPpred e SPpred, descritti nella prima parte del presente elaborato, gli oggetti proposti per questa fase di analisi sono basati su metodi di machine-learning. In particolare SNPs&GO è un tool basato su Support Vector Machine (SVM) che utilizza informazioni derivanti dalla sequenza proteica, dal profilo di sequenza (ovvero da informazioni filogenetiche) e dalle annotazioni funzionali (ottenute dalla banca dati Gene Ontology). Combinando questi tre elementi il predittore è in grado di valutare se una sostituzione amminoacidica nella sequenza proteica di interesse è "Disease Related" o "Neutral". Il predittore PROVEAN (acronimo da Protein Variation Effect Analyzer) svolge la stessa funzione di SNPs&GO, ovvero effettua un'analisi volta a determinare se una mutazione puntiforme ha un impatto sulla funzionalità biologica della proteina. L'algoritmo adottato da PROVEAN si basa sugli allineamenti di sequenza e sull'analisi della conservazione evolutiva dei residui considerati. Diversamente, il tool predittivo I-Mutant è ideato per effettuare una valutazione sulla variazione di stabilità della proteina in seguito ad una sostituzione ammonoacidica. Il predittore è basato su SVM e utilizza come input le informazioni sulla sequenza, sulla posizione del residuo mutato, sul pH e sulla temperatura. Per ciascuna variante analizzata siamo in grado, grazie al risultato fornito dai tre predittori, di dare una indicazione sull'effetto che la mutazione può avere sulla proteina sia a livello di stabilità della struttura proteica sia come possibile associazione alla patologia.

RISULTATI

Il workflow di analisi elaborato del corso dello svolgimento del Dottorato di Ricerca e illustrato nella presente trattazione è stato applicato a numerosi studi sul cancro e ad anche alcuni progetti di ricerca in campo non oncologico. Nella Tabella 1 viene riportata la tipologia e la numerosità dei campioni analizzati.

Tabella 1. Campioni analizzati per mezzo della piattaforma HiScanSQ

Tipologia	# campioni	# reads	GB sequenziate	COVERAGE media
LAM-infant	29(8)	8 X 10^9 (8 X 10^7)	597 (6)	36 X(14-36)
PDAC	20(10)			
GIST	17(14)			
Mioepitelioma	6(2)			
EMCS	6(2)			
DSFP	4(1)			
ACALASIA	8			
CCSK	8			
Sarcomi di Ewing	16			
Liposarcoma	3			
Ganglioneuroma	1			
Ipereosinofilia	1			
Fibromatosi	1			

Nella colonna #campioni i numeri tra parenti fanno riferimento al numero di pazienti per i quali è stata analizzata anche la controparte normale, nella colonna #reads e GB sequenziate i numeri tra parentesi si riferiscono al numero medio di reads per campione e di basi sequenziate, nella colonna COVERAGE media tra parentesi vengono indicati i valori minimo e massimo della Coverage media.

In particolare il contributo bioinformatico è stato dato nei progetti che vengono sommariamente illustrati di seguito:

- Identificazione di varianti a singolo nucleotide e riarrangiamenti cromosomici in pazienti pediatrici affetti da Leucemia Mieloide Acuta (LAM-infant) a cariotipo normale;
- Caratterizzazione genomica di pazienti affetti da Adenocarcinoma Duttale Pancreatico (PDAC)

- Analisi NGS di Tumori Stromali Gastrointestinali (GIST) metastatici con mutazione nell'esone 11 del gene KIT;
- Analisi NGS di GIST con mutazione p.D824V nel gene PDGFRA;
- Analisi NGS di GIST wild type;
- Identificazione di varianti in Liposarcomi;
- Caratterizzazione genomica di pazienti pediatrici affetti da Sarcomi a Cellule Chiare del Rene (CCSK);
- Progetti in collaborazione con l'Istituto Nazionale Tumoti di Milano che ha visto la caratterizzazione genomica di differenti tipologie di sarcomi tra cui: Darmatofibrosarcomi Protuberans (DFSP), Condromiosarcomi Extrascheletrici (EMSC), Mioepiteliomi.
- Progetto in collaborazione con la divisione di Genetica Medica del Policlinico Sant'Orsola – Malpighi inerente lo studio di espressione genica e di identificazione di varianti in pazienti affetti da Acalasia;
- Progetto in collaborazione con Istituto Rizzoli di Bologna inerente la caratterizzazione genomica di Sarcomi di Edwing.

Nella appendice si riporta la lista delle pubblicazioni e la versione integrale lavori pubblicati in riviste scientifiche internazionali.

Case Study

Nella sezione precedente si è fatta menzione dei numerosi progetti di ricerca che sono stati condotti per mezzo di indagini NGS e per i quali è stata applicata la pipeline di analisi bioinformatica qui presentata. Tra quelli citati, uno dei progetti a cui si è dedicata maggiore attenzione è stato quello inerente la caratterizzazione genetica di un sottogruppo raro di GIST cosiddetti wild-type (GIST-WT). La denominazione GIST-WT deriva dal fatto che questa categoria di pazienti (circa il 10-15% del totale) non presenta le mutazioni comunemente note nei recettori delle tirosin-chinasi KIT o PDGFRA.
Per questi tumori si è condotto uno studio volto all'identificazione delle lesioni genetiche causative nonché di nuovi potenziali target terapeutici[50]. Tale studio, che ha avuto grande risonanza tra gli esperti mondiali di GIST, ha permesso di individuare mutazioni a livello della subunità A della Succinico Deidrogenasi (SDHA). Ricerche successive hanno dimostrato che i geni del complesso SDH possono essere considerati marcatori comuni in una percentuale significativa dei GIST-WT, ed è con grande interesse che si sta valutando tale complesso in ottica farmacogenomica.
Nel progetto, oltre all'applicazione del wok-flow di analisi dati, è stato dato il supporto bioinformatico per l'analisi funzionale a livello proteico delle mutazioni messe in evidenza mediante NGS. Va infatti

sottolineato che le variazioni trovate nei geni, siano esse SNV, Indel, riarrangamenti cromosomici o livelli di espressione alterati, vanno in ultima battuta osservate dal punto di vista proteico, in quanto la proteina è l'effettiva responsabile dell'alterazione riscontrabile nella cellula tumorale.

Sono stati analizzati due pazienti giovani adulti (età alla diagnosi 28 e 30 anni) affetti da GIST-WT nei quali vengono evidenziate mutazioni a livello della SDHA. L'analisi con NGS e la successiva validazione per mezzo di Sanger sequencing hanno individuato la mutazione nonsenso p.S384X omozigote nel tumore, la stessa variante è stata trovata come evento eterozigote nel sangue periferico (che rappresenta la controparte normale). Nel secondo paziente viene riscontrata la mutazione somatica missenso p.R589W ed una mutazione germinale nonsenso p.R31X. Ne risulta, quindi, che in entrambi i pazienti c'è un primo evento germinale eterozigote nel gene della SDHA ed un secondo evento (perdità dell'eterozigosità o nuova mutazione) che fa perdere la funzionalità di entrambi gli alleli (Figura 4). A partire da questi risultati sono stati analizzati altri 14 GIST-WT a sono state trovate 4 nuove SNV in 2 pazienti. Globalmente quindi sono state identificate 7 mutazioni in 4 pazienti su 16 (25%). L'analisi condotta con i tool bioinformatici classifica tutte le varianti trovate come destabilizzanti la struttura proteica e potenzialmente associate alla patologia (Tabella 2). Un ulteriore analisi effettuata con l'applicativo Ligplot[51] mostra per ciascuna sostituzione una alterazione evidente delle interazioni non covalenti nella proteina in cui è presente variante.

Tabella 2. Valutazione dell'effetto delle SNV sulla proteina SDHA con i tools SNPs&GO, Provean e I-Mutant

Mutation	SNPs&GO predizione	SNPs&GO reliability index	Provean predizione	Provean score	I-Mutant predizione	I-Mutant reliability index	I-Mutant ΔΔG (kcal/mol)
p.R589W	Disease	9	Deleterious	-6.985	Decrease	2	-0,22
p.R589Q	Disease	9	Deleterious	-3.502	Decrease	6	-0,72
p.R171C	Disease	9	Deleterious	-7.596	Decrease	3	-0,87
p.G419R	Disease	9	Deleterious	-7.022	Decrease	2	-0,37
p.E564K	Disease	5	Deleterious	-3.514	Decrease	8	-0,61

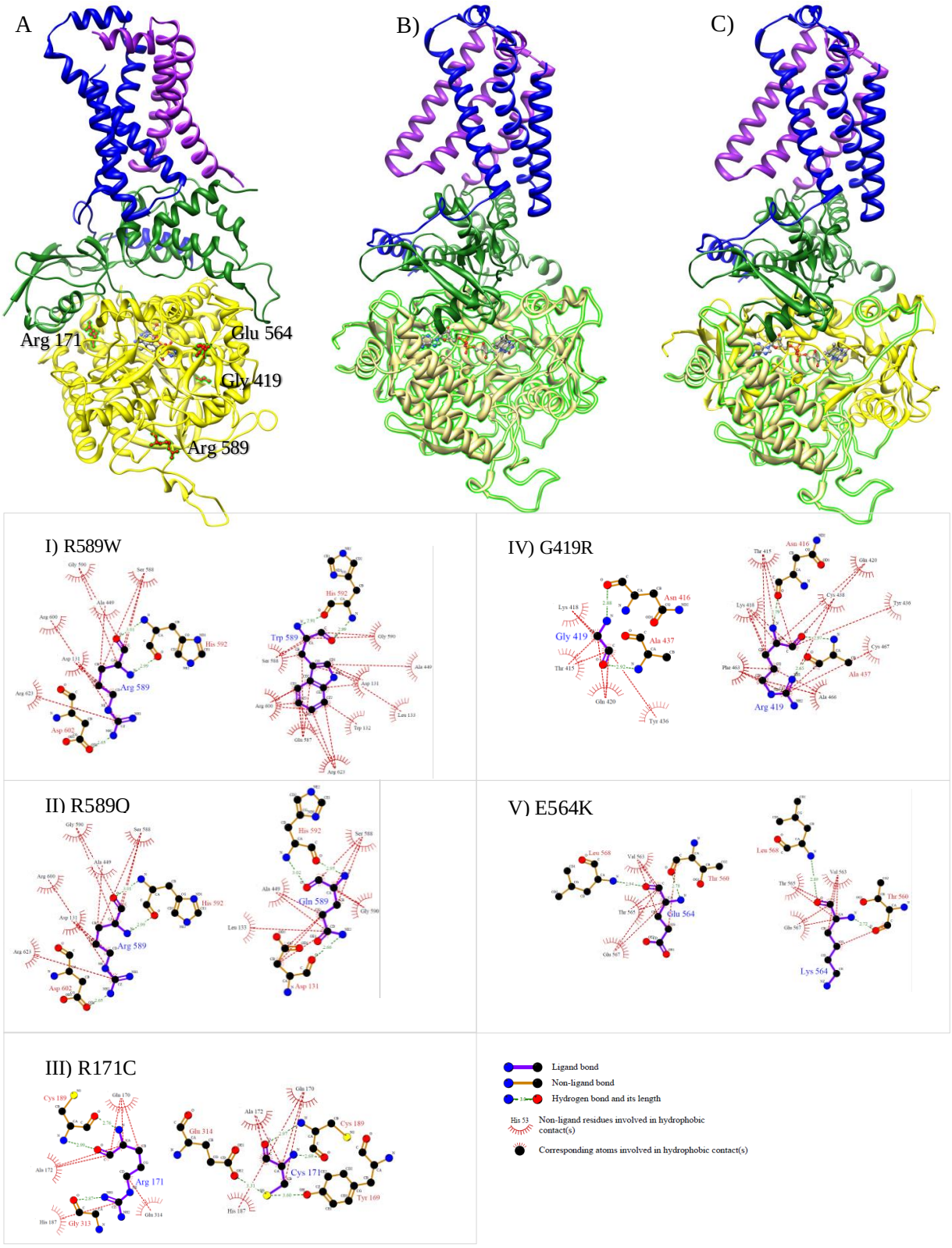


Figura 4. Mutazioni identificate in SDHA in 4 pazienti affetti da GIST-WT. I primi due pazienti analizzati con NGS presentano uno la variante missenso p.R589W (A-I) più la troncazione p.R31X (B) e uno la troncazione omozigote p.S384X (C). Altri due nuovi pazienti analizzati con Sanger sequencing mostrano varianti sul gene di SDHA: uno presenta le SNV p.R589Q e p.R171C (A-II e A-III), mentre le SNV p.G419 e p.E564K sono presenti un altro paziente (A-IV e A-V**).**

CONCLUSIONI

Le tecnologie di sequenziamento massivo rappresentano una vera e propria rivoluzione nel campo della biologia molecolare. In particolare nelle applicazioni oncologiche le metodologie NGS hanno dato, nel corso dell'ultimo decennio, un grande slancio agli approcci di ricerca biomedica improntate nell'ottica della medicina personalizzata. Per mezzo di esperimenti basati su NGS siamo in grado di generare una enorme mole di dati che necessitano di essere interpretati perché possano essere resi fruibili all'intera comunità scientifica. In questo contesto la bioinformatica ha assunto un ruolo cruciale e sempre più essenziale per l'analisi e la gestione dei dati prodotti e, soprattutto, per lo sviluppo di strumenti computazionali, in particolare predittivi, per mezzo dei quali è possibile attribuire un significato biologico al dato di sequenza e di contribuire al consolidamento dell'informazione ancora parziale contenuta nelle banche dati biologiche.

Nel lavoro di ricerca svolto durante il periodo di Dottorato sono stati sviluppati ed applicati metodi computazionali per l'annotazione di proteomi e di genomi. In particolare, nell'ambito degli studi sul cancro condotti per mezzo di NGS, è stata messa a punto una pipeline di analisi di dati avente come obbiettivo finale l'identificazione di geni che possano fungere da possibili marcatori diagnostici o prognostici o da target terapeutici. A questo scopo sono state adottate metodologie bioinformatiche basate sugli approcci predittivi descritti nell'elaborato.

Il workflow di analisi è stato applicato a numerosi progetti di ricerca con risultati notevoli che hanno già ottenuto, specie in alcuni casi, ampio riconoscimento a livello internazionale.

BIBLIOGRAFIA

1. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev. 1996;13:19-50
2. Fleming KG. Riding the wave: structural and energetic principles of helical membrane proteins. CurrOpinBiotechnol. 2000 Feb;11(1):67-71
3. Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schönfisch B,Perschil I, Chacinska A, Guiard B, Rehling P, Pfanner N, Meisinger C. The proteome of Saccharomyces cerevisiae mitochondria. Proc Natl Acad Sci U S A. 2003 Nov 11;100(23):13207-12
4. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000 Jul 21;300(4):1005-16.
5. Dancourt J, Barlowe C. Protein sorting receptors in the early secretory pathway. Annu Rev Biochem. 2010;79:777-802
6. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. 1999 May;8(5):978-84
7. Schein AI, Kissinger JC, Ungar LH. Chloroplast transit peptide prediction: a peek inside the black box. Nucleic Acids Res. 2001 Aug 15;29(16):E82.
8. Claros MG, Vincens P. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem. 1996 Nov 1;241(3):779-86
9. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2(4):953-71
10. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics. 2002 Feb;18(2):298-305
11. Small I, Peeters N, Legeai F, Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics. 2004 Jun;4(6):581-90
12. Petsalaki EI, Bagos PG, Litou ZI, Hamodrakas SJ. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. Genomics Proteomics Bioinformatics. 2006 Feb;4(1):48-55
13. Viklund H, Bernsel A, Skwark M, Elofsson A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics. 2008 Dec 15;24(24):2928-9
14. Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol. 2008 Nov;4(11):e1000213
15. Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004 May 14;338(5):1027-36
16. Käll L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics. 2005 Jun;21 Suppl 1:i251-7
17. Frank K, Sippl MJ. High-performance signal peptide prediction based on sequence alignment techniques. Bioinformatics. 2008 Oct 1;24(19):2172-6

18. Chou KC, Shen HB. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. BiochemBiophys Res Commun. 2007 Jun 8;357(3):633-40.
19. Hiller K, Grote A, Scheer M, Münch R, Jahn D. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W375-9
20. Shen HB, Chou KC. Ensemble classifier for protein fold pattern recognition. Bioinformatics. 2006 Jul 15;22(14):1717-22. Epub 2006 May 3.
21. Fariselli P, Finocchiaro G, Casadio R. SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. Bioinformatics. 2003 Dec 12;19(18):2498-9
22. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011 Sep 29;8(10):785-6
23. Fariselli P, Savojardo C, Martelli PL, Casadio R. Grammatical-Restrained Hidden Conditional Random Fields for Bioinformatics applications. Algorithms Mol Biol. 2009 Oct 22;4:1
24. Li MH, Lin L, Wang XL, Liu T. Protein-protein interaction site prediction based on conditional random fields. Bioinformatics. 2007 Mar 1;23(5):597-604.
25. Culotta A, Kulp D, McCallum A. Gene Prediction with Conditional Random Fields, Technical Report, UM-CS-2005-028, University of Massachusetts, Amherst, 2005.
26. DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE. Conrad: gene prediction using conditional random fields. Genome Res. 2007 Sep;17(9):1389-98. Epub 2007 Aug 9.
27. Sato K, Sakakibara Y. RNA secondary structural alignment with conditional random fields. Bioinformatics. 2005 Sep 1;21 Suppl 2:ii237-42.
28. Lukov L, Chawla S, Church WB. Conditional Random Fields for Transmembrane Helix Prediction. Advances in Knowledge Discovery and Data Mining, 3518, 2005, 155-161.
29. Savojardo C, Fariselli P, Casadio R. BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. Bioinformatics. 2013 Feb 15;29(4):504-5.
30. Vincent M, Passerini A, Labbé M, Frasconi P. A simplified approach to disulfide connectivity prediction from protein sequences. BMC Bioinformatics. 2008 Jan 14;9:20.
31. Savojardo C, Fariselli P, Martelli PL, Shukla P, Casadio R. Prediction of the bonding state of cysteine residues in proteins with machine-learning methods", in R. Rizzo and P.J.G. Lisboa (Eds.): Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2010), LNBI 6685, pp. 98-111, Springer-Verlag Berlin Heidelberg, 2011.
32. Indio V, Martelli PL, Savojardo C, Fariselli P, Casadio R. The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields. Bioinformatics. 2013 Apr 15;29(8):981-8
33. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc. ICML01. 2001. 282-289
34. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ,Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA 3rd, Venter JC. The minimal gene complement of Mycoplasma genitalium. Science.1995 Oct 20;270(5235):397-403

35. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276-7
36. https://support.illumina.com/sequencing/sequencing_software/casava/documentation.ilmn
37. http://hannonlab.cshl.edu/fastx_toolkit/index.html
38. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11
39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9
41. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics.2010 Jan 1;26(1):139-40
42. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. 2011 May;7(5):e1001138
43. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4
44. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics. 2011 Jul 15;27(14):1922-8
45. Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 2010 Mar 15;26(6):730-6
46. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164
47. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009 Aug;30(8):1237-44
48. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688
49. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W306-10
50. Pantaleo MA, Astolfi A, Indio V, Moore R, Thiessen N, Heinrich MC, Gnocchi C, Santini D, Catena F, Formica S, Martelli PL, Casadio R, Pession A, Biasco G. SDHA loss-of-function mutations in KIT-PDGFRA wild-type gastrointestinal stromal tumors identified by massively parallel sequencing. J Natl Cancer Inst. 2011 Jun 22; 103(12):983-7
51. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. 1995 Feb; 8(2):127-34

APPENDICE

LISTA DELLE PUBBLICAZIONI

Full paper in riviste internazionali

- Masetti R, Togni M, Astolfi A, Pigazzi M, Manara E, Indio V, Rizzari C, Rutella S, Basso G, Pession A, Locatelli F. DHH-RHEBL1 fusion transcript: a novelrecurrent feature in the new landscape of pediatric CBFA2T3-GLIS2-positive acute myeloidleukemia. Oncotarget. 2013 Oct;4(10):1712-20.

- Pantaleo MA, Astolfi A, Urbini M, Nannini M, Paterini P, Indio V, Saponara M, Formica S, Ceccarelli C, Casadio R, Rossi G, Bertolini F, Santini D, Pirini MG, Fiorentino M, Basso U, Biasco G. Analysis of allsubunits, SDHA, SDHB, SDHC, SDHD, of the succinatedehydrogenasecomplex in KIT/PDGFRA wild-type GIST. Eur J Hum Genet. 2013 Apr 24.

- Indio V, Martelli PL, Savojardo C, Fariselli P, Casadio R. The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields. Bioinformatics. 2013 Apr 15;29(8):981-8. *

- Masetti R, Pigazzi M, Togni M, Astolfi A, Indio V, Manara E, Casadio R, Pession A, Basso G, Locatelli F. CBFA2T3-GLIS2 fusion transcript is a novel common feature in pediatric, cytogenetically normal AML, not restricted to FAB M7 subtype. Blood. 2013 Apr 25;121(17):3469-72.

- Nannini M, Astolfi A, Paterini P, Urbini M, Santini D, Catena F, Indio V, Casadio R, Pinna AD, Biasco G, Pantaleo MA. Expression of IGF-1 receptor in KIT/PDGF receptor-α wild-type gastrointestinal stromal tumors with succinate dehydrogenase complex dysfunction. Future Oncol. 2013 Jan;9(1):121-6.

- Pantaleo MA, Nannini M, Astolfi A, Biasco G; GIST Study Group Bologna. A distinct pediatric-type gastrointestinal stromal tumor in adults: potential role of succinate dehydrogenase subunit A mutations. Am J SurgPathol. 2011 Nov;35(11):1750-2.

- Pierleoni A, Indio V, Savojardo C, Fariselli P, Martelli PL, Casadio R. MemPype: a pipeline for the annotation of eukaryotic membrane proteins. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W375-80. *

- Pantaleo MA, Astolfi A, Indio V, Moore R, Thiessen N, Heinrich MC, Gnocchi C, Santini D, Catena F, Formica S, Martelli PL, Casadio R, Pession A, Biasco G. SDHA loss-of-function mutations in KIT-PDGFRA wild-type gastrointestinal stromal tumors identified by massively parallel sequencing. J Natl Cancer Inst. 2011 Jun 22;103(12):983-7. *

Nota: Gli articoli contrassegnato con il carattere “*” vengono riportati nella versione integrale

Abstract presentati in convegni nazionali ed internazionali

- CBFA2T3-GLIS2: un nuovo trascritto di fusione identificato mediante sequenziamento massivo del trascrittoma, ricorrente nelle lam pediatriche a cariotipo normale e associato a prognosi sfavorevole.
R. Masetti, M. Pigazzi, M. Togni, A. Astolfi, V. Indio, E. Manara, R. Casadio, A. Pession, G. Basso, F. Locatelli
Congresso AIEOP 09-11/06/2013 Roma

- DHH-RHEBL1: un nuovo trascritto di fusione, identificato mediante whole-transcriptome sequencing, ricorrente nello scenario delle leucemie acute mieloidi infantili CBFA2T3-GLIS2-positive.
M. Togni, R. Masetti, A. Astolfi, M. Pigazzi, V. Indio, G. Biasco, C. Rizzari, G. Basso, F. Locatelli, A. Pession
Congresso AIEOP 09-11/06/2013 Roma

- Identification of novel genetic alterations by whole-transcriptome sequencing in infants with cytogenetically normal acute myeloid leukemia.
Marco Togni , Riccardo Masetti , Annalisa Astolfi , Valentina Indio , Martina Pigazzi, Elena Manara , Rita Casadio , Giuseppe Basso , Andrea Pession , Franco Locatelli
European Hematology Association 18th congress 13-16/06/2013 Stockholm

- Gene fusions evidence in a KIT/PDGFRA wild-type GIST without mutations in SDH units identified by a whole transcriptome study.
Milena Urbini, Annalisa Astolfi, Valentina Indio, Maristella Saponara, Margherita Nannini, Cristian Lolli, Anna Mandrioli, Lidia Gatto, Maria Caterina Pallotti, Guido Biasco, Maria A. Pantaleo
ASCO Annual Meeting 2013 – American Society of Clinical Oncology

- Genome study of PDGFRA D842V mutant GIST using next generation sequencing approach.
Maria A. Pantaleo, Annalisa Astolfi, Milena Urbini, Valentina Indio, Margherita Nannini, Maristella Saponara, Cristian Lolli, Maria Caterina Pallotti, Anna Mandrioli, Lidia Gatto, Guido Biasco
ASCO Annual Meeting 2013 – American Society of Clinical Oncology

- Whole-transcriptome paired-end sequencing and the pancreatic cancer genetic landscape.
Marina Macchini, Annalisa Astolfi, Valentina Indio, Silvia Vecchiarelli, Elisa Grassi, Carla Serra, Riccardo Casadei, Donatella Santini, Marielda D'Ambra, Claudio Ricci, Francesco Minni, Guido Biasco, Mariacristina Di Marco
ASCO Annual Meeting 2013 – American Society of Clinical Oncology

- Identification of single nucleotide variants in gastrointestinal stromal tumor KIT/PDGRFA wild type (WT GIST) with massively parallel sequencing
Indio Valentina, Tasco Gianluca, Martelli Pier L, Casadio Rita, Pantaleo Maria A, Astolfi Annalisa, Formica Serena, Paterini Paola, Bisco Giudo
56^ National Meeting of the Italian Society of Biochemistry and Molecular Biology – Chieti September 2012

- The prediction of organelle targeting peptides in eukariotic proteins with Grammatical Restrained Hidden Conditional Random Fields
Martelli Pier Luigi, Indio Valentina, SavojardoCastranse, Fariselli Piero, Casadio Rita
56^ National Meeting of the Italian Society of Biochemistry and Molecular Biology – Chieti September 2012

- Whole genome discovery of genetic alterations in resecteable and advanced pancreatic cancer
Macchini M., Astolfi A., Casadei R., Ricci C., Indio V., Vecchiarelli S., D'Ambra M., Grassi E., Santini D.4, Minni F., Biasco G., Di Marco M.
CongressoNazionale AIOM 2012

- SNP-Array High Resolution Cytogenetic Analysis of Resectable and Advanced Pancreatic Cancer
Marina Macchini, Annalisa Astolfi, Riccardo Casadei, Claudio Ricci, Valentina Indio, Silvia Vecchiarelli, Marielda D'Ambra, Elisa Grassi, Donatella Santini, Carla Serra, Raffaele Pezzilli, Francesco Minni, Guido Biasco, Mariacristina di Marco
AISP - 36th National Congress. Bologna, Italy. October 4-6, 2012

- Massively Parallel Sequencing Analysis of Genetic Alterations Carried by Pancreatic Adenocarcinoma
Marina Macchini, Annalisa Astolfi, Riccardo Casadei, Valentina Indio, Silvia Vecchiarelli, Claudio Ricci, Marielda D'Ambra, Elisa Grassi, Donatella Santini, Carla Serra, Raffaele Pezzilli, Francesco Minni, Guido Biasco, Mariacristina di Marco
AISP - 36th National Congress. Bologna, Italy. October 4-6, 2012

- SDHA and SDHB mutations in KIT/PDGFRA WT gastrointestinal stromal tumors.
Margherita Nannini, Maria A. Pantaleo, Annalisa Astolfi, Milena Urbini, Serena Formica, Valentina Indio, Chiara Gnocchi, Maristella Saponara, Cristian Lolli, Maria Caterina Pallotti, Anna Mandrioli, Lidia Gatto, Alessandra Maleddu, Rita Casadio, Guido Biasco
ASCO Annual Meeting 2012 – American Society of Clinical Oncology

- Identification of protein variants in gastrointestinal stromal tumor KIT/PDGFRA wild type (WT GISTs) with RNA massively parallel sequencing and computational analysis
V. Indio, G. Tasco, P.L. Martelli, R Casadio
M. A. Pantaleo, A. Astolfi, S. Formica, P. Paterini, G. Biasco
36th FEBS CONGRESS 2001 - Federation of the Societies of Biochemistry and Molecular Biology, Torino, Italy

- Identification of Single Nucleotide Variants in Gastrointestinal Stromal Tumor KIT/PDGRFA Wild Type (WT GIST) with Massively Parallel Sequencing
Indio V, Tasco G, Martelli PL, Casadio R, Pantaleo MA, Astolfi A, Formica S, Paterini P, Biasco G
BITS 2011 Annual Meeting of the Bioinformatics Italian Society Pisa, Italy

- Identification of single nucleotide variants in gastrointestinal stromal tumor KIT/PDGFRA wild type (WT GISTs) by massively parallel sequencing
V. Indio, M. A. Pantaleo, A. Astolfi, R. Casadio, P. Paterini, S. Formica, P. Martelli, R. Moore, N. Thiessen, M. di Battista, F. Catena, D. Santini, M. C. Heinrich, C. Gnocchi, A. P. Dei Tos, G. Biasco;
ASCO Annual Meeting 2011 – American Society of Clinical Oncology

- Identification of SDHA (subunit A of the succinate dehydrogenase) mutations in KIT/PDGFRA WT gastrointestinal stromal tumors (GISTs).
M.A. Pantaleo, A. Astolfi, V. Indio, P. Paterini, S. Formica, R. Casadio, P. Martelli, A. Maleddu, M. Nannini, A. P. Dei Tos, M. C. Heinrich, D. Santini, F. Catena, C. Ceccarelli, M. Fiorentino, M. di Battista, R. Moore, N. Thiessen, C. Gnocchi, G. Biasco;
ASCO Annual Meeting 2011 – American Society of Clinical Oncology

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 8 2013, pages 981–988
doi:10.1093/bioinformatics/btt089

Sequence analysis

Advance Access publication February 21, 2013

The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields

Valentina Indio[1,2], Pier Luigi Martelli[1,3,*], Castrense Savojardo[1,4], Piero Fariselli[1,4] and Rita Casadio[1,2,3]

[1]Biocomputing Group, University of Bologna, 40126 Bologna, [2]Giorgio Prodi Interdepartmental Center for Cancer Research, University of Bologna, 40138 Bologna, [3]Department of Biology and [4]Department of Computer Science and Engineering, University of Bologna, 40126 Bologna, Italy

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Targeting peptides are the most important signal controlling the import of nuclear encoded proteins into mitochondria and plastids. In the lack of experimental information, their prediction is an essential step when proteomes are annotated for inferring both the localization and the sequence of mature proteins.

Results: We developed TPpred a new predictor of organelle-targeting peptides based on Grammatical-Restrained Hidden Conditional Random Fields. TPpred is trained on a non-redundant dataset of proteins where the presence of a target peptide was experimentally validated, comprising 297 sequences. When tested on the 297 positive and some other 8010 negative examples, TPpred outperformed available methods in both accuracy and Matthews correlation index (96% and 0.58, respectively). Given its very low–false-positive rate (3.0%), TPpred is, therefore, well suited for large-scale analyses at the proteome level. We predicted that from ~4 to 9% of the sequences of human, *Arabidopsis thaliana* and yeast proteomes contain targeting peptides and are, therefore, likely to be localized in mitochondria and plastids. TPpred predictions correlate to a good extent with the experimental annotation of the subcellular localization, when available. TPpred was also trained and tested to predict the cleavage site of the organelle-targeting peptide: on this task, the average error of TPpred on mitochondrial and plastidic proteins is 7 and 15 residues, respectively. This value is lower than the error reported by other methods currently available.

Availability: The TPpred datasets are available at http://biocomp.unibo.it/~valentina/TPpred/. TPpred is available on request from the authors.

Contact: gigi@biocomp.unibo.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 30, 2012; revised on January 15, 2013; accepted on February 18, 2013

1 INTRODUCTION

Mitochondria and plastids (in plants) are membrane-enclosed organelles contained in eukaryotic cells that take part into essential biological processes, including cell bioenergetics and metabolism. Following the endosymbiotic theory, the DNA-containing organelles derive from free-living bacteria that have been incorporated into the cytoplasm of early eukaryotic cells. The organelles retain a small portion of their original genome, encoding few tens and few hundreds proteins in mitochondria and plastids, respectively. However, both experimental and computational methods estimate that some thousands different proteins are located in organelles (Sickmann *et al.*, 2003; van Wijk, 2004). The last release of MitoMiner lists 2755 human proteins experimentally characterized as mitochondrial with mass spectrometry or with green-fluorescent protein tagging (Smith *et al.*, 2012), and the AT_CHLORO database lists 1323 *Arabidopsis thaliana* proteins annotated as chloroplastic on the basis of mass spectrometry experiments (Ferro *et al.*, 2010). Most organellar proteins are, therefore, encoded by the nuclear genome, synthesized by the cytoplasmic ribosomes and then targeted to the organelles.

The most common, although not unique, type of targeting signal consists of an N-terminal sequence (the targeting peptide, often referred as pre-sequence in mitochondria and transit peptide in plastids). The targeting peptide is cleaved when the native protein is translocated through the outer membrane of the organelles by means of the translocon protein complexes. In multicellular organisms, it has been estimated that ~10–25% of nuclear genes encode for proteins endowed with an N-terminal peptide targeting the protein to mitochondria and plastids (Emanuelsson *et al.*, 2000). Targeting peptides are highly heterogeneous in terms of length (ranging from 10 to 150 residues) and primary sequence (Bruce, 2001; Patron and Waller, 2007; Staiger *et al.*, 2009). Phylogenetic and structural studies recognized a modular architecture in targeting peptides, often consisting of two or three separate domains, potentially forming amphiphilic α-helices or β-strands when interacting with the organelle outer membrane (Bruce, 2001; Habib *et al.*, 2007). The modular organization is probably involved in the sub-organellar trafficking of proteins and gives origin to a great variety of primary sequences for targeting peptides (Texeira and Glazer, 2012). In a small but increasing number of proteins, the same targeting peptide can mediate the translocation to both the mitochondria and the chloroplasts (Carrie *et al.*, 2009).

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

Presently, in UniProtKB, only 11% of ~9200 sequences that are endowed with a targeting peptide are supported by experimental annotations. Because of the biological importance of targeting mechanisms and the scarcity of experimentally validated knowledge, machine-learning tools have been introduced to predict the presence of targeting peptides and the position of the corresponding cleavage sites. Early methods are specific for chloroplasts, e.g. ChloroP (Emanuelsson *et al*., 1999) and PCLR (Schein *et al*., 2001), and mithochondria, e.g. MitoProt (Claros and Vincens, 1996). More recent tools integrate the prediction of targeting peptides with the prediction of secretory signal peptides, although they have different compositional and length features. Among these are TargetP (Emanuelsson *et al*., 2007), which incorporates ChloroP, iPSORT (Bannai *et al*., 2002), Predotar (Small *et al*., 2004) and PredSL (Petsalaki *et al*., 2006).

All methods predict the presence of the targeting peptide, but only MitoProt, TargetP and PredSL predict the position of the cleavage site. In general, all the predictors analyse an N-terminal portion of the native protein, ranging from 40 to 100 residues, depending on the method. ChloroP (Emanuelsson *et al*., 1999), TargetP (Emanuelsson *et al*., 2007), PCLR (Schein *et al*., 2001) and Predotar (Small *et al*., 2004) are based on neural networks (NNs) and their input includes residue composition, hydrophobicity and abundance of charged residues. iPSORT adopts a rule-based algorithm that considers 434 different propensity scales (Bannai *et al*., 2002). MitoProt defines a discriminant function on a pool of 47 physicochemical properties (Claros and Vincens, 1996). PredSL combines NNs, hidden Markov models (HMM) and scoring matrices (Petsalaki *et al*., 2006). Because of the paucity of data, all methods have been trained on datasets containing also non-experimentally validated targeting peptides, predicted with computational methods and/or inferred by similarity.

Prediction of targeting peptide can be considered as a labelling problem, where residues of the N-terminal region of the sequence are assigned either to 't' (targeting peptide) or 'n' (non-targeting peptide) labels. Grammatical-Restrained Hidden Conditional Random Fields (GRHCRF) is a recently introduced machine-learning tool well suited to solve labelling problems (Fariselli *et al*., 2009, Savojardo *et al*., 2011). GRHCRFs offer several advantages: (i) like HMMs, they can incorporate previous knowledge on the problem by introducing a grammar on the prediction labels; (ii) like the Hidden Conditional Random Fields (HCRF), they are discriminative models and do not require the strong independence assumptions made in HMMs (Fariselli *et al*., 2009; Lafferty *et al*., 2001); and (iii) similar to NNs, they can analyse complex and heterogeneous input encodings.

Here, we introduce TPpred, a new predictor for targeting peptides based on GRHCRFs. TPpred is trained on a non-redundant dataset containing only experimentally validated targeting peptides and efficiently predicts both the presence of targeting peptides and the localization of the cleavage sites.

2 METHODS

2.1 Dataset

We gathered the eukaryotic proteins longer than 45 residues from SwissProt (release November 2011) and annotated with existing evidence at the protein level, with the exclusion of fragments. Starting from this set, we collected both the positive and the negative datasets. The positive dataset (proteins endowed with an experimentally detected targeting peptide) was collected searching in the feature field for the keyword 'TRANSIT PEPTIDE', which identifies all the pre-sequences directing a protein to an organelle in UniProtKB (http://www.uniprot.org/keywords/KW-0809). We excluded annotations labelled as 'by similarity', 'probable' or 'potential', and we retained only proteins from mitochondria and plastids provided with a known cleavage site. Proteins lacking the keyword 'TRANSIT PEPTIDE' were collected in the negative set. By this, we obtained 757 positive and 47 363 negative examples. To obtain a non-redundant dataset, sequences were then compared with Basic Local Alignment Search Tool, and a graph was built linking the pairs of sequences (nodes) that share >30% identity on local alignments (HSP) with e-value $<10^{-3}$. The graph was clustered by extracting its connected components with a transitive closure algorithm. After this procedure, sequences in different clusters share <30% identity. We also checked that the 160-residue long N-terminal regions of sequences are <30% identical when extracted from different clusters. The non-redundant training dataset was built by randomly selecting one sequence per cluster. The final dataset consists of 297 sequences with targeting peptide (DB+) and 8010 without targeting peptide (DB−) (Table 1). To test whether the prediction is affected by the presence of transmembrane helices, we extracted a subset of proteins with an α-helix annotated in the 160 residue-long N-terminal segment by UniProtKB (values in parentheses of Table 1). The dataset is available at: http://biocomp.unibo.it/~valentina/TPpred/.

2.2 GRHCRF

Prediction of targeting peptides can be posed as a labelling problem with a strong grammatical constraint: the targeting peptide region ('t') precedes the non-targeting peptide region ('n'). Starting from HMMs that are the prototypical models addressing this type of problems, Conditional Random Fields (CRF) have been introduced: they are discriminative models that allow relaxing the strong independence assumptions of HMMs by means of a global normalization procedure (Lafferty *et al*., 2001). GRHCRFs have been developed to overcome the limitations of the coincidence between labels and states typical of CRFs (Fariselli *et al*., 2009; Savojardo *et al*., 2011). GRHCRFs decouple the set of labels from the set of states and allow defining a one-to-many mapping between them. Like HMMs, GRHCRFs can be represented through an automaton comprising a set of labelled states connected by transitions. The topology of the automaton casts the grammar to be modelled. The same label can be shared among different states. This ensures a great expressive power of the method and a large flexibility in the automaton design. A feature function is associated to each state and to each transition. The parameters of the feature functions are learned from the association between the input sequences included in the training set and their known labellings. Discriminative learning has been implemented for finding the parameters that maximize the probability of a label given the input (see Fariselli *et al*., 2009 for details). Given a trained model, the labelling of a

Table 1. The training dataset

Organism	Without TP (DB−)	Chloroplastic TP	Mitochondrial TP	With TP (DB+)
Plants	605 (86)	95 (12)	18 (0)	113 (12)
Non-plants	7405 (1081)	—	184 (12)	184 (12)
Total	8010 (1167)	95 (12)	202 (12)	297 (24)

TP, Targeting peptide. Values in parentheses refer to proteins where a transmembrane helix is annotated in the 160 residue-long N-terminal segment.

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

sequence is predicted with the posterior-Viterbi algorithm. This implements a decoding procedure that preserves the grammar, and it is based on the posterior probabilities for each label as computed by the model (Fariselli *et al.*, 2005).

2.3 Input features

For each sequence, 160 N-terminal residues were considered for building the input to GRHCRFs. Each position of the segment was encoded with a 25-valued vector describing the type and the physicochemical features of the corresponding residue. The 25-valued vector comprises four different modules: (i) a 20-valued binary vector, describing the residue type, whose elements are all null but the one corresponding to the residue to be encoded (seq); (ii) one value encoding the average Kyte–Doolittle hydrophobicity (Kyte and Doolittle, 1982) of a seven-residue long window centred on the residue to be encoded (kd); (iii) two values encoding the number of positively and negatively charged residues in a seven-residue window (ch); and (iv) two values describing the hydrophobic moments (hm) computed considering 100° and 160° angles for simulating ideal α-helices and β-sheets, respectively. The program *hmoment* included in EMBOSS (Rice *et al.*, 2000) was adopted to carry out the computation of the hydrophobic moments. For each position in the sequence, the feature function of each GRHCRF state considers an 11-residue window; therefore, it takes in input $11 \times 25 = 275$ different variables. When encoding the five N- and C-terminal residues of each sequence, we padded the empty positions of the window with '0' values.

2.4 Training procedure

We adopted a 5-fold cross-validation procedure for training and testing, by randomly splitting the non-redundant training set into five subsets. Three subsets were used for training the method (training set), one for validation (validation set) and the remaining for evaluating the performance (test set). The best model topology, the best parameters and the best input were selected on the basis of the results obtained on the validation set. Five training runs were performed, and performance was computed collecting all the results obtained for the five test sets.

2.5 Scoring the performance

Different scoring indexes were used to evaluate the prediction performances at the protein level. For the two protein classes, namely, 'with targeting peptide' (+) and 'without targeting peptide' (−), we indicated with TP and TN the number of true-positive and true-negative predictions, respectively, and with FP and FN the number of false-positive and false-negative predictions, respectively.

General prediction scores are the overall accuracy (Acc) and the Matthews correlation coefficient (MCC), defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}} \tag{2}$$

Unlike Acc, MCC is only slightly affected by bias deriving from the unbalance between positive and negative examples in the dataset, and it is then the most useful global score. Scoring indexes evaluating the performance on the single class are sensitivity (Sn), specificity (Sp) and false-positive rate (FPR) computed as:

$$Sn(c) = \frac{TP}{TP + FN} \tag{3}$$

$$Sp(c) = \frac{TP}{TP + FP} \tag{4}$$

$$FPR(c) = \frac{FP}{TN + FP} \tag{5}$$

where c is the class at hand. A thorough explanation of the purposes of these indexes can be found in Baldi *et al.* (2000).

2.6 Prediction with available methods

For sake of comparison, we predicted the sequences included in our dataset with the following methods: (i) the TargetP server was accessed at http://www.cbs.dtu.dk/services/TargetP/; (ii) the executable version of iPSORT was downloaded from http://ipsort.hgc.jp/caml-iPSORT/; (iii) the executable version of PredSL was downloaded from http://hannibal.biol.uoa.gr/PredSL/source.html; (iv) the Predotar server was accessed at http://urgi.versailles.inra.fr/predotar/predotar.html; (v) the software of MitoProt was downloaded from ftp://ftp.biologie.ens.fr/pub/molbio; and (vi) PCLR was re-implemented in house using the parameters listed in http://www.andrewschein.com/cgi-bin/pclr/weights.html. When required, the proper prediction parameters were selected, dividing plant and non-plant proteins.

2.7 Whole-proteome analysis

The complete sets of proteins from *Homo sapiens* (GRHh37.p5), *A.thaliana* (TAIR10) and *Saccharomyces cerevisiae* (EF4) were downloaded from the EnsEMBL website (www.ensembl.org). These sets comprise 93 588, 35 386 and 6692 protein sequences (including splicing variants), respectively. They are encoded by 21 160, 27 416 and 6692 genes, respectively. We predicted all the proteins with our TPpred and checked the agreement between the prediction of targeting peptide and the Gene Ontology (GO) annotation of the subcellular localization as reported in EnsEMBL. We retained only experimental annotations labelled with the following evidence codes: EXP (experimental), IDA (inferred from direct assay), IPI (inferred from physical interaction), IMP (inferred from mutant phenotype), IGI (inferred from genetic interaction) and IEP (inferred from expression pattern) (http://www.geneontology.org/GO.evidence.shtml).

3 RESULTS AND DISCUSSION

3.2 Main features of targeting peptides

3.1.1 Length The length of the mitochondrial and plastidic targeting peptides ranges from 10 up to 150 residues (Fig. 1). Mitochondrial-targeting peptides are, on average, shorter than plastidic ones, being the average lengths 35 and 59 residues, respectively. The dispersions around the average lengths are, however, very high and of ∼16 and 22 residues, respectively. We chose not to separate the two datasets, because of (i) the scarcity of non-redundant proteins experimentally annotated for targeting peptides and (ii) the possibility that the same targeting peptide mediates the translocation to both mitochondria and plastids (Carrie *et al.*, 2009).

3.1.2 Residue composition The residue composition of targeting peptides is plotted in Figure 2 and compared with the whole-sequence composition of the proteins included in our dataset. Relative standard deviations of the samples are not shown for sake of clarity and have been evaluated to be ∼20% of the plotted data. Mitochondrial and plastidic proteins (represented in Fig. 2 with blue and cyan bars, respectively) do not show major compositional differences with proteins included in the negative set (red bars). On the contrary, targeting peptides are characterized by a peculiar composition. The differences in composition between the targeting peptides and the whole sequences have been assessed in terms of log-odds and *P*-values

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

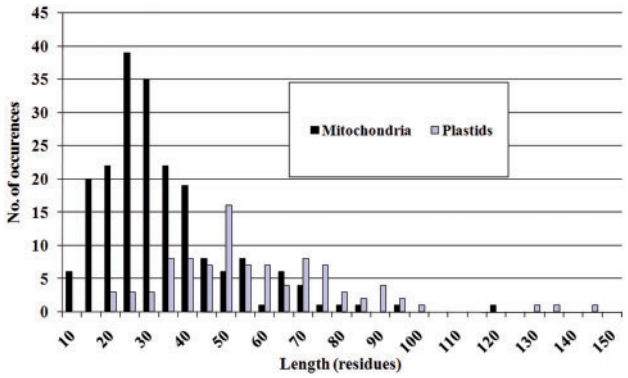


Fig. 1. Length distribution of the targeting peptides of proteins included in the DB+ dataset

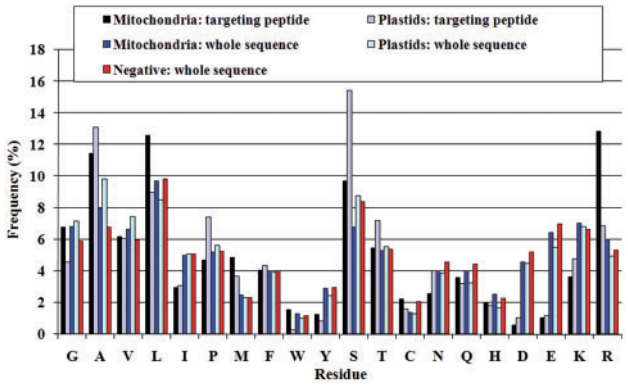


Fig. 2. Residue composition of targeting peptides and of sequences included in the training set. Average values are computed on 202 mitochondrial proteins, 95 plastidic proteins and 8010 proteins non-containing the targeting peptide (negative). Relative standard deviations of the samples, evaluating the dispersion around the reported average value, are ~20%, and they are not represented in the figure

(Supplementary Table S1). At a significance level equal to 10^{-10}, both mitochondrial- and plastidic-targeting peptides (black and grey bars, respectively) are enriched in alanine (A) and serine (S) and depleted in negatively charged residues (D, E), tyrosine (Y) and isoleucine (I). Other differences can be detected when mitochondrial- and plastidic-targeting peptides are separately analysed: the former are enriched in methionine (M), leucine (L) and arginine (R), whereas lysine (K) is underrepresented. In plastidic-targeting peptides glycine (G) is less frequent. The composition of targeting peptides of our dataset is similar to that of previous analyses (Texeira and Glaser, 2012), and it accounts for the interactions with proteins involved in protein import and peptide cleavage, whose structural details are still unknown (Jarvis and Robinson, 2004; Pfanner and Geissler, 2001).

3.1.3 Cleavage site The strongest compositional information is thought to reside in the region neighbouring the cleavage site, as it is recognized by the active site of peptidase complexes. Starting from the 297 proteins of the positive dataset, we aligned the eight residues downstream and upstream the cleavage site. The resulting profile is visualized in the sequence logo of Figure 3. The cleavage site is between positions 0 and 1. The logo

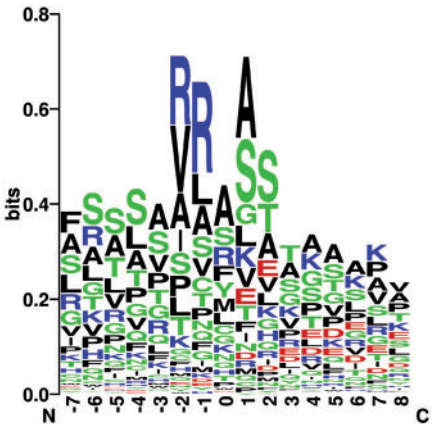


Fig. 3. Sequence logo of the positions neighbouring the cleavage site. Sequence logo (Schneider and Stephens, 1990) is computed by the WebLogo server (weblogo.berkeley.edu). Position '1' is the first residue of the mature protein. Height of letters is proportional to their information content in profile. Information is measured in bits and ranges between 0 and $\log_2(20) \approx 4.3$. Colour codes cluster residues in apolar (black), polar (green), positively charged (blue) and negatively charged (red)

represents the conservation of each residue present in a position with a letter whose height is proportional to its information content, ranging from 0 to $\log_2 20 = 4.3$. It is evident that the positions neighbouring the cleavage site are slightly more conserved than the others, although the overall information content is moderate. The most conserved residues are in position −1 and −2, downstream, and +1 and +2, upstream the cleavage site.

3.2 The automaton for predicting the targeting peptide

We cast the information on the features of targeting peptides in the automaton described in Figure 4. It comprises 45 states associated to the label 't' (targeting states represented with squares) and 31 states associated to the label 'n' (non-targeting states, represented with circles). The overall model mainly consists of two sub-models: the grey coloured states describe the 160 N-terminal residues of proteins with targeting peptides and white coloured states describe the same portion in proteins without targeting peptide.

The sub-model for proteins without targeting peptide consists of states connected with a dense topology that allows a general description of a broad range of different sequences.

The sub-model for proteins endowed with targeting peptides is more specific and consists of different groups of states aiming at capturing the modular organization of targeting peptides. As reported in Section 1, phylogenetic and structural studies recognized a modular architecture in targeting peptides that often contain two or three separate domains, potentially forming amphiphilic α-helices or β-strands when interacting with the organelle outer membrane (Bruce, 2001; Habib *et al.*, 2007). The targeting model consists of four modules: (i) the N-terminal region, probably unstructured and variable in length, is modelled with a set of densely connected states; (ii) two central domains, consisting of forward-connected states, aim at capturing the features of the modules described in literature; and (iii) the cleavage

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

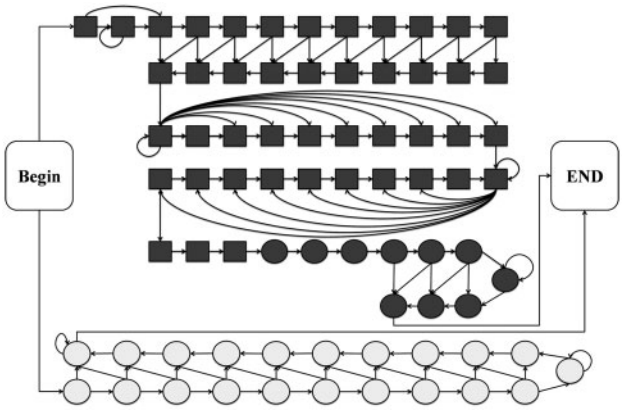


Fig. 4. Automaton for targeting peptide prediction. Squares represent states labelled as targeting peptide ('t'), whereas circles represent non-targeting peptide states ('n'). Grey states model the 160-residue long N-terminal region of proteins endowed with targeting peptide (target model). White states model proteins devoid of signal peptide (non-target model). See text for further details

site region, where three residues upstream and three residues downstream are explicitly modelled with six different states. To model the large variability of targeting peptide lengths, several states are self-connected. Owing to the overall topology, the length of predicted targeting peptides spans from 10 to 155 residues, in agreement with the lengths deduced from the analysis of our dataset (Fig. 1). As the modular organization cannot be easily recognized in all the known targeting peptides, the model topology has to maintain a high flexibility.

Different model topologies have been adopted, and we retained the model best performing on the validation sets. The performance scores are computed in cross-validation on test sets independent of both the training and the validation sets.

3.3 Targeting peptide prediction with different inputs

The GRHCRF was trained adopting the strategy described in Section 2.4 and adopting a 5-fold cross-validation procedure. Three sets are used for training, one (validation set) for choosing the best automaton parameters (including the window size and the input encoding) and the remaining (testing set) for computing the indexes scoring the performance. As the complete dataset was reduced for similarity (Section 2.1), this procedure ensures a reliable evaluation of the generalization capability of the method.

Table 2 lists the performance scores evaluated on the complete set (altogether comprising 297 positive and 8010 negative examples) when different inputs are fed to the GRHCRF. The single sequence leads to an overall accuracy value (Acc) as high as 95% and an MCC value equal to 0.50. By incrementally adding the different features described in Section 2.3, performance increases on testing sets and reaches the maximum value when information on sequence, hydrophobicity, charge and hydrophobic moment of the N-terminal regions are included. The final performance of TPpred is as high as 96% accuracy and 0.58 MCC. When proteins belonging to organisms from different kingdoms are evaluated separately, the MCCs are as high as 0.74 and 0.52 for plants and non-plants, respectively (for a more detailed evaluation see Supplementary Table S2).

Table 2. Performance of GRHCRF with different input encoding

Input	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(−) (%)	Sp(−) (%)	FPR (%)
Seq	95	0.5	69	40	96	99	3.8
Seq + kd	95	0.54	73	42	96	99	3.7
Seq + kd + ch	96	0.56	75	46	97	99	3.3
Seq + kd + ch + hm (TPpred)	96	0.58	75	48	97	99	3.0

seq, sequence; kd, Kyte–Doolittle hydropathy; ch, charge; hm, hydrophobic moment. For a thorough description of the input, see Section 2.3. The evaluation dataset comprises 297 positive and 8010 negative examples. Scoring indexes are computed with a 5-fold cross-validation procedure, by collecting the results on the test sets. For index definition, see Section 2.4.

To assess the advantage in adopting GRHCRF for predicting the presence of targeting peptides, we also implemented a predictor based on an HMM with the same topology described in Figure 4 and trained on the same dataset. When sequence information is adopted as input, the HMM scores with an MCC equal to 0.39, significantly lower than that reached with GRHCRF on the same input ('seq' line in Table 2; see Supplementary Table S3 for a detailed comparison).

3.4 Benchmark with available methods

Table 3 lists the prediction performance of TPpred on the non-redundant dataset as compared with that of other available methods predicting both mitochondrial- and plastidic-targeting peptides. TPpred outperforms all of these methods. When evaluated with the Fisher r-to-z transformation (Fisher, 1921), the difference in MCC between TPpred and the best performing predictor (Predotar) is significant, with $P < 10^{-4}$. It is worth noticing that in Table 3, only TPpred is evaluated by adopting a cross-validation procedure. Indeed, the overlap between the training datasets of other methods and that adopted for training/validating TPpred can lead to overestimate the performances of the other tools.

The low specificity on the positive class [Sp(+)] is the major pitfall of all available predictors, probably because of the unbalance in the datasets adopted for training them. TPpred is by far the most specific predictor (48%). This is reflected in a lower sensitivity, that, however, reaches a high value [Sn(+) = 75%]. Moreover, TPpred scores with an FPR (3%) that is less than a half with respect to the best available tools. The same trend is confirmed, when comparing with predictors specific for mitochondria (MitoProt) and plastids (PCLR) (Table 4): also in this case, TPpred scores with the highest accuracy and MCC.

TPpred scores with a lower sensitivity than other methods and this corresponds to an increase of the false-negative rate [FNR = 1 − Sn(+)]. FNR is equal to ~25% for TPpred, higher than that reported by other methods (7–21%). However, when implementing a prediction method suitable for large-scale annotation of proteins, it is important to keep the error rate as low as possible in the most abundant class, to keep the number of wrong predictions low. At the whole-proteome level, proteins without targeting peptide (the negative set) are by far more

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

Table 3. Benchmark results on non–organelle-specific predictors

Method	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(−) (%)	Sp(−) (%)	FPR (%)	FPR TM[a] (%)
TPpred[b]	96	0.58	75	48	97	99	3	2.5
TargetP[c]	89	0.44	93	24	89	100	10.9	11.1
Predotar[c]	92	0.49	90	30	92	100	7.7	8.8
iPSORT[c]	89	0.38	79	22	90	99	10.2	12.4
PredSL[c]	91	0.45	88	26	91	100	9.4	12.5

Scoring indexes are computes as described in Section 2.4. Tools other than TPpred were run as described in Section 2.5. The evaluation dataset comprises 297 positive and 8010 negative examples. [a]FPR computed on the negative examples endowed with a transmembrane helix within the 160 N-terminal residues (see Table 1, first column, values in parentheses). [b]Sequences are predicted in cross-validation on the test sets. [c]The benchmark dataset overlaps with the training set.

Table 4. Benchmark results on organelle-specific predictors

Method	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(−) (%)	Sp(−) (%)	FPR (%)	FPR TM[a] (%)
Mitochondrial proteins								
202 positive (mitochondrial) and 8010 negative(1167 TM) examples from all eukaryotes								
TPpred[b]	96	0.52	74	39	97	100	3	2.5
MitoProt[c]	77	0.23	89	9	76	100	22	29.9
Plastidic proteins								
95 positive (plastidic) and 605 negative (86 TM) examples from plants								
TPpred[b]	94	0.73	73	81	97	96	2.6	0
PCLR[c]	86	0.6	93	48	84	99	15.5	12.8

Scoring indexes are computes as described in Section 2.4. Tools other than TPpred were run as described in Section 2.5. Only the values relative to TPpred are computed in cross-validation. [a]FPR computed on the negative examples endowed with a transmembrane helix within the 160 N-terminal residues (see Table 1, first column, values in parentheses). [b]Sequences are predicted in cross-validation on the test sets. [c]The benchmark dataset overlaps with the training set.

abundant than proteins with targeting peptide (the positive set), and this is why we consider that the low FPR (that is the rate of error on the negative set) reported by TPpred is an interesting feature for its adoption in large-scale analyses.

We also tested the FPRs of predictors on the subset of 1167 negative examples endowed with a transmembrane helix within the 160 N-terminal residues (last column of Tables 3 and 4). Less than 2.5% of this set is predicted by TPpred as proteins endowed with a targeting peptide, and this is the best result obtained with the currently available methods. This enables TPpred to be safely adopted for analysing membrane proteins. In particular, when used as a pre-filter to identify cleaved peptides, it lowers the risk of removing N-terminal transmembrane helices.

3.5 Prediction of the cleavage site

TPpred also predicts the position of the cleavage site along the sequence. This information is important, as it allows knowing the sequence of the mature and functional protein. Some predictors, however, do not report it (e.g. Predotar). In Table 5, we comparatively assessed the prediction of cleavage site performed with TPpred, TargetP, PredSL and MitoProt (with the last one, only on mitochondrial proteins). Mitochondrial and plastidic sequences are evaluated separately because of the different average length of the corresponding targeting peptides (35 and 59 residues, respectively, see also Section 3.1.1). For each prediction, we evaluated the error (E) as the difference between the positions of the real and the predicted cleavage sites. We then computed the mean error (ME) and the number of prediction for which the error is lower than the standard deviation (σ) of the length distribution of targeting peptides ($E < \sigma$ score). Standard deviations are equal to 16 and 22 residues for mitochondria and plastids, respectively, as discussed in Section 3.1.1. Our TPpred outperforms the other methods both in terms of ME and $E < \sigma$ score, particularly for mitochondria. This indicates that TPpred correctly predicts the correct length of the targeting peptide, even if the length distribution is spread.

3.6 Prediction of targeting peptides in whole proteomes

The whole proteomes of three species were downloaded from EnsEMBL and predicted with TPpred. Results of the prediction are reported in Table 6. We estimate that 4.0, 9.0 and 6.1% of proteins are endowed with targeting peptide in human, *Arabidopsis* and yeast, respectively. The estimates are somewhat lower than those previously reported with other methods (10–25%, Emanuelsson *et al.*, 2000). This result is possibly because of the low FPR of TPpred (3%) that limits the number of mispredictions in the negative set.

For proteins predicted with targeting peptide, we tested the compatibility with the GO annotations for cellular component reported in EnsEMBL and labelled with an experimental evidence code (see Section 2.6). GO terms were divided into three subsets: (i) terms directly related to mitochondrial or plastidic

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

Table 5. Benchmark on the cleavage site prediction

Method	Mitochondria		Plastids	
	ME (res)	E<σ score (%)	ME (res)	E<σ score (%)
TPpred[a]	7	89	15	74
TargetP[b]	12	71	16	71
PredSL[b]	12	75	17	73
MitoProt[b]	13	75	—	—

ME: mean prediction error on the position of the cleavage site. E<σ score: proportion of predictions with error lower than the standard deviation of the length distribution of targeting peptides. [a]Sequences are predicted in cross-validation on the test sets. [b]The benchmark dataset overlaps with the training set.

Table 6. Targeting peptides predicted at the whole-organism scale

	H.sapiens	*A.thaliana*	*S.cerevisiae*
Whole organism			
No. of proteins (no. of genes)	93 588 (21 160)	35 386 (27 416)	6692 (6692)
With predicted targeting peptide			
No. of proteins (no. of genes)	3744 (1685)	3194 (2521)	407 (407)

Predicted proteomes are available at http://biocomp.unibo.it/~valentina/TPpred/. Values reported in parentheses refer to the number of genes.

Table 7. Comparison between targeting peptide predictions and experimental GO annotations

GO annotation (EnsEMBL)	*H.sapiens*	*A.thaliana*	*S.cerevisiae*
Mitochondrion	288 (8%)	286 (9%)	228 (56%)
Plastid	—	1297 (41%)	—
Compatible	5 (0%)	10 (0%)	3 (1%)
Incompatible	158 (4%)	221 (7%)	40 (10%)
Not annotated	3293 (88%)	1370 (43%)	136 (33%)

Percentage values are computed with respect to the number of protein sequences predicted as endowed with targeting peptide (3744 in *Homo*, 3194 in *Arabidopsis* and 407 in *Saccharomyces*).

localizations; (ii) terms compatible with mitochondrial or plastidic localizations, as they include them as subsets (e.g. cell part, intracellular, cytoplasm and membrane); and (iii) terms incompatible with mitochondrial or plastidic localization.

The results for the three proteomes are reported in Table 7. In the case of the human proteome, only 12% of the proteins predicted with targeting peptide are endowed with experimental annotation of their localization: 8% of proteins are mitochondrial and are, therefore, correctly predicted; 4% are localized in other cellular components (mostly in the nucleus) and can, therefore, be considered as false predictions. The rate of experimental annotation in *A.thaliana* and yeast is much higher (57 and 67%, respectively) and mostly confirms the predictions of TPpred: 50 and 56% of proteins predicted with targeting peptide in *Arabidopsis* and yeast, respectively, are localized in mitochondria or plastids. When considering only the set of annotated proteins, the rates of success in the two well-annotated organisms are, therefore, 50/57 = 87% and 56/67 = 84%, respectively. Proteins with incompatible localization are, in both organisms, mostly annotated as nuclear or, in the case of *Arabidopsis*, as plasma membrane.

The good agreement between the prediction and the experimental annotations confirms the suitability of TPpred for performing prediction of whole proteomes. In the three analysed organisms, we also suggest a new annotation for a large amount of proteins: 3293 in human, 1370 in *Arabidopsis* and 136 in yeast.

4 CONCLUSIONS

We implemented TPpred, a new predictor for targeting peptides in mitochondrial and plastidic proteins. TPpred is based on GRHCRFs, a recently introduced machine-learning approach. Differently from available methods, it is trained only on experimentally validated targeting peptides. TPpred outperforms other available methods, both in predicting the presence and the length of targeting peptides. TPpred is significantly more specific than the available predictors and scores with a very low-FPR. This feature makes TPpred useful for predicting the targeting peptides in proteomes of whole organisms. In particular, when tested on the proteomes of *H.sapiens*, *A.thaliana* and *S.cerevisiae*, the estimate of the amount of proteins endowed with targeting peptide is ~4–9%. The good agreement between the predictions of TPpred and the experimental annotations suggests that this method can be combined with subcellular localization predictors for improving their performance in genome-wide annotation procedures.

Funding: This work has been supported by the following grants: PRIN 2009 project 009WXT45Y (Italian Ministry for University and Research: MIUR), COST BMBS Action TD1101 (European Union RTD Framework Program) and PON project PON01_02249 (Italian Ministry for University and Research: MIUR). PhD fellowship of the Italian Ministry for University and Research: MIUR (to C.S.).

Conflict of Interest: none declared.

REFERENCES

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bannai,H. *et al.* (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.

Bruce,B.D. (2001) The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta*, **1541**, 2–21.

Carrie,C. *et al.* (2009) Protein transport in organelles: dual targeting of proteins to mitochondria and chloroplasts. *FEBS J.*, **276**, 1187–1195.

Claros,M.G. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

Emanuelsson,O. *et al.* (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.*, **8**, 978–984.
Emanuelsson,O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
Emanuelsson,O. *et al.* (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat. Protoc.*, **2**, 953–971.
Fariselli,P. *et al.* (2005) A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics*, **6**, S12.
Fariselli,P. *et al.* (2009) Grammatical-restrained hidden conditional random fields for bioinformatics applications. *Algorithms Mol. Biol.*, **4**, 13.
Ferro,M. *et al.* (2010) AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol. Cell. Proteomics*, **9**, 1063–1084.
Fisher,R.A. (1921) On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3–32.
Jarvis,P. and Robinson,C. (2004) Mechanisms of protein import and routing in chloroplasts. *Curr. Biol.*, **14**, R1064–R1077.
Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
Lafferty,J. *et al.* (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc. ICML01*, 282–289.
Habib,S.J. *et al.* (2007) Analysis and prediction of mitochondrial targeting signals. *Methods Cell Biol.*, **80**, 761–781.
Patron,N.J. and Waller,R.F. (2007) Transit peptide diversity and divergence: a global analysis of plastid targeting signals. *BioEssays*, **29**, 1048–1058.
Petsalaki,E.I. *et al.* (2006) PredSL: a tool for the N-terminal sequence-based prediction of subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.
Pfanner,N. and Geissler,A. (2001) Versatility of the mitochondrial protein import machinery. *Nat. Rev. Mol. Cell Biol.*, **2**, 339–349.
Rice,P. *et al.* (2000) EMBOSS: European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
Savojardo,C. *et al.* (2011) Improving the prediction of disulfide bonds in eukaryotes with machine learning methods and protein subcellular localization. *Bioinformatics*, **27**, 2224–2230.
Schein,A.I. *et al.* (2001) Chloroplast transit peptide prediction: a peek inside the black box. *Nucleic Acids Res.*, **29**, e82.
Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
Sickmann,A. *et al.* (2003) The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl. Acad. Sci. USA*, **103**, 13207–13212.
Small,I. *et al.* (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
Smith,A.C. *et al.* (2012) MitoMiner: a data warehouse for mitochondrial proteomics data. *Nucleic Acids Res.*, **40**, 1060–1067.
Staiger,C. *et al.* (2009) Diversity in degrees of freedom of mitochondrial transit peptides. *Mol. Biol. Evol.*, **26**, 1773–1780.
Texeira,P.F. and Glaser,E. (2012) Processing peptidases in mitochondria and chloroplasts. *Biochim. Biophys. Acta.*, **1833**, 360–370.
van Wijk,K.J. (2004) Plastid proteomics. *Plant Physiol. Biochem.*, **42**, 963–977.

Downloaded from http://bioinformatics.oxfordjournals.org/ by Pier Luigi Martelli on March 25, 2014

Published online 4 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W375–W380***
doi:10.1093/nar/gkr282

MemPype: a pipeline for the annotation of eukaryotic membrane proteins

Andrea Pierleoni[1,*], Valentina Indio[2,3], Castrense Savojardo[2], Piero Fariselli[2], Pier Luigi Martelli[2] and Rita Casadio[2,3]

[1]Externautics s.p.a., Externautics s.p.a. – Bioinformatics, Via Fiorentina 1, 53100 Siena, [2]Bologna Biocomputing Group, Bologna Computational Biology Network, University of Bologna and [3]Interdepartmental Center for Cancer Research 'Giorgio Prodi' (CIRC), University of Bologna, Italy

Received February 25, 2011; Revised April 1, 2011; Accepted April 12, 2011

Downloaded from http://nar.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

ABSTRACT

MemPype is a Python-based pipeline including previously published methods for the prediction of signal peptides (SPEP), glycophosphatidylinositol (GPI) anchors (PredGPI), all-alpha membrane topology (ENSEMBLE), and a recent method (MemLoci) that specifically discriminates the localization of eukaryotic membrane proteins in: 'cell membrane', 'internal membranes', 'organelle membranes'. MemLoci scores with accuracy of 70% and generalized correlation coefficient (GCC) of 0.50 on a rigorous homology-unbiased validation set and overpasses other predictors for subcellular localization. The annotation process is based both on inheritance through homology and computational methods. Each submitted protein first retrieves, when available, up to 25 similar proteins (with sequence identity $\geq$50% and alignment coverage $\geq$50% on both sequences). This helps the identification of membrane-associated proteins and detailed localization tags. Each protein is also filtered for the presence of a GPI anchor [0.8% false positive rate (FPR)]. A positive score of GPI anchor prediction labels the sequence as exposed to 'Cell surface'. Concomitantly the sequence is analysed for the presence of a signal peptide and classified with MemLoci into one of three discriminated classes. Finally the sequence is filtered for predicting its putative all-alpha protein membrane topology (FPR $<$1%). The web server is available at: http://mu2py.biocomp.unibo.it/mempype.

INTRODUCTION

In Eukaryotes, most protein functional features are constrained by the different cell compartments and their enclosing membranes (1–3). Functional features of biological membranes strictly depend on proteins that specifically interact with them. Membrane proteins can be classified into two major classes: integral membrane proteins, which span the lipid bilayer [transmembrane (TM) proteins (TPs)] or covalently bind a lipid molecule, and peripheral membrane proteins, which physically interact with the membrane surfaces. About 30% of eukaryotic proteins in SwissProt are annotated with the keyword 'membrane' (48 963 sequences out of 166 219), and 75% of them are also annotated as 'transmembrane' (37 659 sequences). In most cases, the experimental determination of the structure and function of membrane proteins is presently hampered by technical problems and their function is often annotated on the basis of sequence similarity. Our annotation procedure takes advantage of both inheritance of annotation (annotation transfer) after homology search and annotation by predicting features with different machine learning approaches. To this purpose MemPype integrates methods that are specifically suited to predict the presence of signal peptides, lipid anchors, membrane protein localization and topology of all-alpha membrane proteins, thus providing an integrated computational resource for annotation of eukaryotic membrane proteins. However, the main novelty in MemPype is the integration of MemLoci, a method that allows a reliable classification of both eukaryotic integral and peripheral membrane proteins into three classes: cell membrane (CM), organelle membranes (OMs) and internal membranes (IMs) (4). This is a key step for functional annotation of membrane proteins in relation to their membrane type (5,6). We propose MemPype to support annotation of membrane proteomes of eukaryotic organisms with the unique feature of also identifying proteins present on the cell surface. These chains are likely candidates to be characterized as biomarkers and/or targets for new drugs.

*To whom correspondence should be addressed. Tel: +39 0577 231275; Fax: +39 0577 43444; Email: andrea.pierleoni@externautics.com; andrea@biocomp.unibo.it

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

MEMPYPE WORKFLOW

MemPype includes two flows of annotation (Figure 1). The first collects information directly from SwissProt in terms of keywords and Gene Ontology (GO) terms associated with proteins sharing high similarity with the target sequence (≥50% sequence identity with an alignment coverage ≥50% on both sequences, see below). The second parallel flow of annotation includes machine learning-based methods that score at the state of the art for the specific problem at hand. Each sequence is filtered for the presence of: (i) signal peptides with SPEP (7); (ii) presence and location of glycophosphatidylinositol (GPI)-anchoring domains with PredGPI (8); then (iii) the subcellular localization of both integral and peripheral membrane proteins is predicted with MemLoci, a recent predictor based on support vector machine (SVM); and finally (iv) the location and topology of all-alpha integral membrane proteins is predicted with ENSEMBLE 3.0 (9). The only input is the residue sequence of the target protein. The first step of the pipeline is a BLAST search against SwissProt that produces alignments of the target sequence with an E-value $\leq 10^{-3}$ (leftmost path in Figure 1). Homologous sequences are used both for performing annotation transfer by sequence similarity and for compiling the sequence profiles that are used as input to most of the predictive methods included in the pipeline (rightmost path in Figure 1). Both flow outputs are given as a result of MemPype running (Figure 2). The results of the first search gives at the most 25 aligned sequences and their features as derived from SwissProt. This information can or cannot be present depending on the target sequence. The second output is always present and gives computed features whose reliability is statistically computed according to the different predictors and can be inspected in relation to the results of the SwissProt search when available. The platform integrates predictors that have been previously described and validated on their specific task. Presently a set of proteins with experimentally validated features to be used in cross-validation for the joint combination of all the predictors is not available. Prediction performances are therefore calculated independently for each method with never seen before

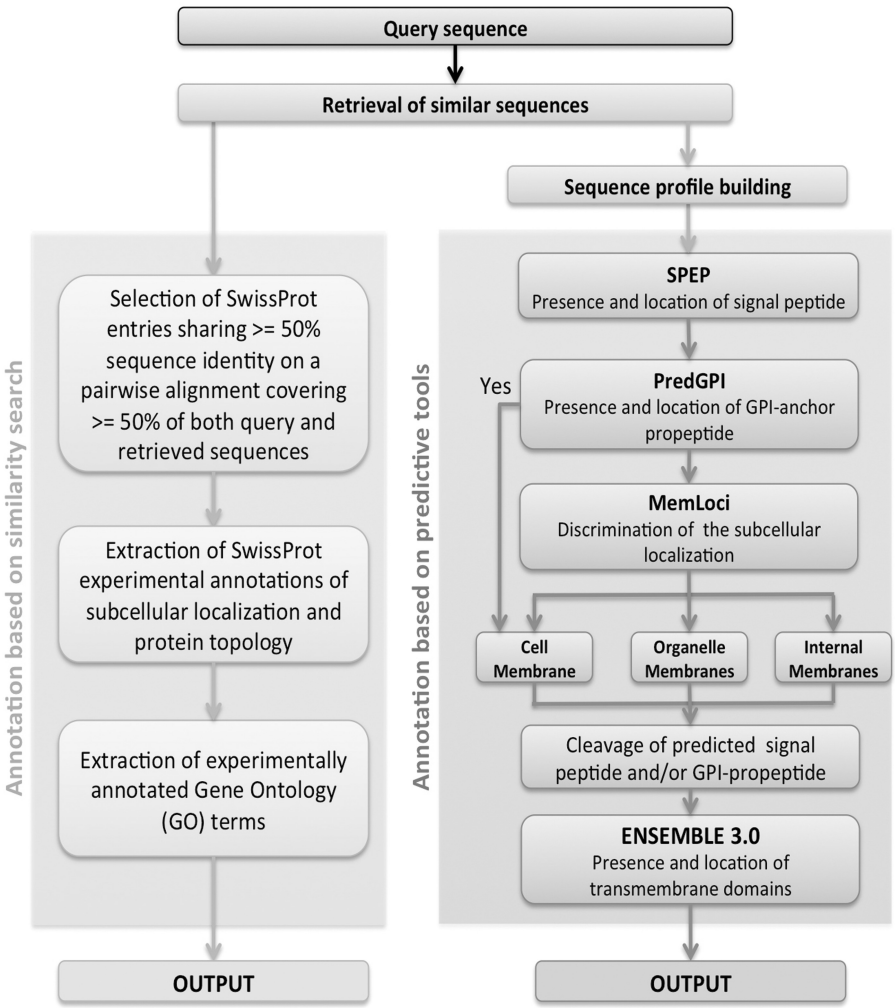
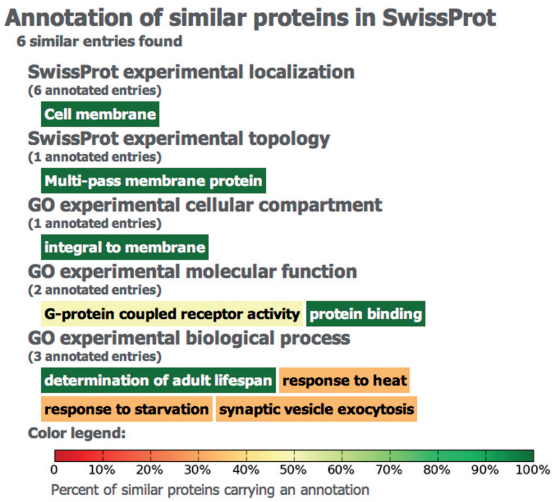


Figure 1. Workflow of the MemPype annotation pipeline. MemPype performs annotation with homology search and prediction tools. See text for further details.

Downloaded from http://nar.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014



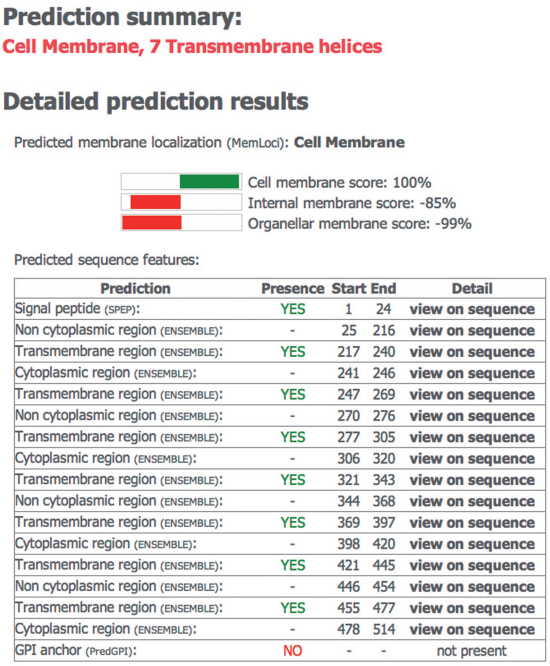


Figure 2. MemPype output results. Two outputs are returned: (i) a list of at the most 25 proteins sharing sequence identity ≥50% on an alignment covering ≥50% of both sequence lengths (when available). Both keywords and GO terms can be transferred on the basis of sequence similarity to the query sequence. (ii) A list of all the predicted features including signal peptide [with SPEP (7)], GPI-anchor [with PredGPI (8)], all-alpha TM topology [with ENSEMBLE3.0 (9)] and prediction of subcellular localization [with MemLoci (4)]. See text for further details.

proteins carrying along the experimentally validated property to be predicted.

ANNOTATION THROUGH INHERITANCE

Transfer of annotation on the basis of sequence similarity is a widely adopted procedure that relies on the assumption that similar sequences share similar structural and functional features (10). The threshold value of sequence similarity necessary for ensuring a reliable inference of function depends on the specific task. It is well known that the overall protein structure is conserved for proteins sharing some ≥30% identical residues, while the conservation of molecular function requires higher identity thresholds [≥50% (11)]. In relation to subcellular localization, sequence identity ≥30% ensures a reliable annotation transfer within non-membrane proteins (12). However, to our knowledge, the same threshold has not yet been determined for membrane proteins. To this aim, we collected from SwissProt 24 640 membrane proteins endowed with experimental annotation of subcellular localization [the set is described in (4)]. Twelve localization classes are considered. Upon an extensive pairwise alignment procedure, we determined that the subcellular localization is conserved in 99.7% cases, when two proteins share ≥50% sequence identity with coverage ≥50% on both sequences (data not shown). The MemPype annotation transfer procedure considers therefore only the set of annotated SwissProt sequences fulfilling these constraints with respect to the target proteins. When many annotated sequences with identity ≥50% and coverage ≥50% are retrieved, only the most similar 25 are taken into account. If existing, the annotations reported in the 'KEYWORD' field of the retrieved sequences and referring to structural and localization features are collected, as well as the GO annotations coming from experimental evidences. All the annotation terms are then represented as a tag cloud, where each tag is coloured with a scale representing the frequency of each keyword in the set (Figure 2). By pointing over each tag, the detailed statistics of each annotation appears. The set of entries promoting a specific annotation can then be retrieved by clicking on the corresponding tag. In some cases, the annotation transfer procedure allows a very specific and detailed annotation such as 'Endoplasmic reticulum-Golgi intermediate compartment membrane.' Moreover, the system can be useful for annotating proteins endowed with multiple localizations. It is not always possible to find annotated proteins fulfilling the constraints of sequence identity necessary for a reliable transfer of annotation based on homology search. A complementary approach is therefore the adoption of predictive methods that run in the same platform and whose results can be either compared/confirmed with those obtained with the homology search or provides the unique annotation resource.

PREDICTION OF SIGNAL PEPTIDE AND GPI ANCHOR

The first step of the prediction pipeline is to determine the sequence of the mature protein, where N-terminal signal

Downloaded from http://nar.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

peptides and/or the GPI-anchoring propeptides, when present, are cleaved. To this aim, SPEP in its version for eukaryotic sequences (7) and PredGPI (8) are applied. Both methods analyse the residue sequence and efficiently determine the presence of peptides as well as the position of the cleavage sites. SPEP is a neural network (NN)-based system, trained on 2300 eukaryotic proteins endowed with experimental annotation (13). Two NNs scan the 65-residue long N-terminal segment of the query sequence, scoring the probability of each residue to be part of a signal peptide and to be the cleavage site, respectively. The allowed signal peptide length ranges between 11 and 59 residues. A signal peptide is predicted if the sum of the outputs of the NNs are greater than a threshold that was selected in order to optimize the performance. By this, when performing the discrimination task on the training data set with a cross-validation procedure, SPEP scores with a Matthews correlation coefficient (CC) as high as 0.91 and overall accuracy (Acc) equal to 95% (7). Here a validation set consisting of 1287 eukaryotic proteins has been extracted from (14) with the exclusion of sequences present in the SPEP training set. The results of the blind validation are reported in Table 1 and show a performance consistent with the scores obtained in cross-validation (CC = 0.87 and Acc = 93%). PredGPI is trained on a data set comprising 340 and 10 630 GPI- and non-GPI-anchored proteins, respectively (8). It includes a SVM, whose discrimination threshold is selected in order to limit the false positive rate (FPR) to 0.5% on the training set. By this, the cross-validation performances are CC = 0.78 and Acc = 99% (8). When a protein is predicted as GPI anchored, the cleavage site is predicted with a hidden Markov model (HMM) that casts the features of the cleaved propeptide and its surrounding regions. Here we collect a validation set consisting of 19 GPI-anchored proteins (with unknown cleavage site) released after training PredGPI, and 391 non-GPI-anchored proteins released after Jan 2011. On this blind set PredGPI scores with CC = 0.87 and Acc = 99.2%, with FPR of the GPI-anchored class as low as 0.8% (Table 1). MemPype outputs list, when present, cleaved peptides highlighted along the sequence. Sequence and sequence profile of the mature protein are then obtained by deleting the sequence segments corresponding to the cleaved peptides. When a sequence contains a GPI-anchor domain, its subcellular localization is labelled 'cell membrane' (15). The low FPR of PredGPI ensures that the rate of wrong localization annotation due to misprediction of GPI anchor is about 1%. Irrespective of this labelling, the sequence is predicted by the complete pipeline and results of MemLoci and the possible presence of TM helices are reported (see next sections). To further assess the error rate that could arise from the combination of PredGPI and MemMoci, PredGPI was also scored on a blind validation subset of MemLoci comprising 68 proteins in OM and IM with the exclusion of CM proteins. Only one protein is wrongly predicted as GPI anchored and thus reported as 'cell membrane', confirming the low FPR of PredGPI.

PREDICTION OF SUBCELLULAR LOCALIZATION

Prediction of subcellular localization of eukaryotic membrane proteins is performed with MemLoci [4], a SVM-based method able to discriminate the localization of membrane proteins within three classes: CM, OMs and IMs. The OM class comprises proteins located at mitochondrial or plastidial membranes: the IM class comprises all the remaining intracellular membranes (the endoplasmic reticulum, the nuclear membranes, the Golgi apparatus, the vesicles, the vacuoles, the lysosomes, the peroxisome, the microsomes and the endosome). MemLoci is the first tool specifically suited to predict the subcellular localization of both integral and peripheral membrane proteins. Other available predictors of subcellular localization explicitly exclude membrane proteins from their training sets (16,17), group all the membrane proteins into a single class referred as 'membrane' or 'cell membrane' (18,19), or focus on specific membrane types and organisms (20,21). MemLoci scores with generalized CC (GCC) (22) in the range of 0.50 when tested on both

Table 1. Performance of the different predictors included in MemPype on never seen before validation sets

Method	Blind validation set	Sen, %	Sp, %	FPR, %	Acc, %	CC
SPEP	543 proteins with SP	89	95	3	93	0.87
	744 proteins without SP	97	91	11		
PredGPI[a]	19 GPI-anchored proteins	89	85	0.8	99	0.87
	391 non-GPI-anchored proteins	99	99	11		
ENSEMBLE3.0[a]	15 TM proteins	100	83	0.4	99	0.91
	208 non-TM proteins	99	100	0		
MemLoci[a]	32 CM proteins	56	75	9	70	0.50[b]
	18 OM proteins	50	56	9		
	50 IM proteins[c]	86	72	34		

[a]The validation set collects never seen before chains by the method and deposited after January 2010. Predictions are scored with the following indexes: Sen: sensitivity = (no. of correctly predicted proteins in the class)/(total no. of proteins in the class); Sp: specificity = (no. of correctly predicted proteins in the class)/(total no. of proteins predicted in the class); FPR = (no. of mispredicted proteins in the class)/(total no. of proteins in the complementary class); Acc = (no. of correctly predicted proteins)/(total no. of proteins); Matthews CC is adopted for binary classifications, while GCC ([b]) is computed for multiclass classifications (22).
[c]IMs comprising all the endomembrane system except the cell membrane. All the validation sets are available at the MemPype website in the 'Info' page.

Downloaded from http://nar.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

the 10 634 sequences included in the training set and the 100 sequences of an independent validation set (Table 1). For each sequence, MemPype lists the localizations predicted with MemLoci and three values scoring their likelihood. The highest value indicates the most likely prediction.

TOPOLOGY PREDICTION AND DISCRIMINATION AND OF ALL-ALPHA TPs

The mature sequence (after signal peptide and GPI-anchor propetide cleavage) is predicted for the presence and topology of all-alpha TM domains with ENSEMBLE3.0, an updated version of ENSEMBLE (9) and based on an ensemble prediction of different machine learning tools that analyse the information contained in sequence profiles, including the capability of discriminating between all-alpha membrane and globular protein. ENSEMBLE 3.0 is trained on a non-redundant data set of 138 all-alpha membrane proteins (including only three eukaryotic chains), whose structure is known with atomic resolution and was deposited in the Protein Data Bank (PDB) before January 2010. Performing a rigorous cross-validation, ENSEMBLE3.0 is able to correctly locate the TM segments of 126 proteins (91%) and to predict the correct orientation with respect to the membrane plane of 119 proteins (86%) of the training/testing set, respectively. Here we test ENSEMBLE 3.0 on a validation set of 15 independent membrane proteins sharing low identity ($\leq$25%) with the training set and whose structures have been deposited after January 2010. This set includes only three proteins from eukaryotes, and two of these are endowed with one validated and one putative signal peptide, respectively. When the sequences of all 15 mature proteins are predicted, ENSEMBLE3.0 correctly computes the topology of all of them. Alternatively, when the full-length sequence of the 15 proteins is submitted to ENSEMBLE 3.0, the topology of only 13 proteins is correctly predicted (87%), with the exclusion of the two eukaryotic proteins endowed with signal peptide. These proteins are correctly predicted when SPEP is combined with ENSEMBLE3.0. In order to test whether ENSEMBLE3.0 is capable of discriminating membrane from globular proteins, we trained a filter on a data set also including 1611 globular structural domains, relative to proteins sharing <25% sequence similarity with the training set and released before January 2010 [extracted from PDB with PISCES (23)]. On a validation set comprising 208 never seen before globular domains (in proteins released after January 2010 and with sequence identity $\leq$25% to the training set) and the 15 TM proteins, FPR was 0 and 0.4%, respectively (Table 1). When the total set of eukaryotic full-length globular and membrane proteins (67 and 3, respectively) were jointly predicted by SPEP and ENSEMBLE, FPR was 0 and 2%, respectively. For TPs, MemPype lists the membrane spanning segments and their topological organization (cytoplasmic, non-cytoplasmic; Figure 2). When the sequence does not contain predicted membrane-spanning segments or GPI-anchored domains, a warning message is visualized indicating that MemLoci prediction should be taken with caution and possibly validated by merging features derived from the homology search.

WEB SERVER

The MemPype web server requires protein sequences in FASTA format as input. Each sequence must at least be 50-residue long. Upon request submission the server displays the prediction result page that is periodically updated until the completion of the prediction procedure. This page can be bookmarked and accessed later. Moreover, a unique identifier marks each prediction request as a future reference to retrieve prediction results. For each sequence the current queue state is reported, and upon completion the prediction results are shown. These are stored in a local database and will remain available for at least 1 month. The web server can be accessed either from anonymous or registered users. Registration is free of charge. Registered users can submit up to five sequences per request and up to 30 different requests per hour, while, to enforce a fair use policy, anonymous users are allowed for only 1 sequence per request and 10 requests per hour. For facilitating the retrieval of the results the web server provides a 'Recent Jobs' page, where the predictions of anonymous users are publicly available, while registered users can retrieve their own jobs in the private 'My Jobs' page. All the software used to build MemPype (except for BLAST+) is written in Python language. The web server runs on a web2py engine, and the annotated sequences are stored in SQLite database adopting the BioSQL schema. Parsing of SwissProt annotation data is performed with the BioPython uniprot-xml parser. HMMs and SVMs needed for all the prediction steps were implemented in Python as well.

ACKNOWLEDGEMENTS

C.S. and V.I. are PhD students supported by Ministero Italiano della Università e Ricerca (MIUR) and CIRC, respectively.

FUNDING

MIUR-FIRB (Fondo per gli Investimenti della Ricerca di Base) 2003/LIBI-International Laboratory for Bioinformatics (to R.C., in part). Funding for open access charge: Fondo Ordinario per le Università (FFO) 2010 (to R.C. and P.L.M.).

Conflict of interest statement. None declared.

REFERENCES

1. Sachs,J.N. and Engelman,D.M. (2006) Introduction to the membrane protein reviews: the interplay of structure, dynamics, and environment in membrane protein function. *Annu. Rev. Biochem.*, **75**, 707–712.
2. White,S.H. (2009) Biophysical dissection of membrane proteins. *Nature*, **459**, 344–346.
3. Almén,M.S., Nordström,K.J.V., Friedriksson,R. and Schiöt,H.B. (2009) Mapping the human membrane proteome: a majority of

Downloaded from http://nar.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.*, **7**, 50.
4. Pierleoni,A., Martelli,P.L. and Casadio,R. (2011) MemLoci: predicting subcellular localization of membrane proteins in Eukaryotes. *Bioinformatics*, **27**, 1224–1230.
5. Imai,K. and Nakai,K. (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, **10**, 3970–3983.
6. Casadio,R., Martelli,P.L. and Pierleoni,A. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomics Proteomics*, **7**, 63–73.
7. Fariselli,P., Finocchiaro,G. and Casadio,R. (2003) SPEPLip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**, 2498–2499.
8. Pierleoni,A., Martelli,P.L. and Casadio,R. (2008) PredGPI: a GPI-anchor predictor. *BMC Bioinformatics*, **9**, 392.
9. Martelli,P.L., Fariselli,P. and Casadio,R. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, **19**, i205–i211.
10. Loewenstein,Y., Raimondo,D., Redfern,O.C., Watson,J., Frishman,D., Linial,M., Orengo,C., Thornton,J. and Tramontano,A. (2009) Protein function annotation by homology-based inference. *Genome Biol.*, **10**, 207.
11. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
12. Nair,R. and Rost,B. (2002) Sequence conserved for subcellular localization. *Prot. Sci.*, **11**, 2836–2847.
13. Menne,K.M., Hermjakob,H. and Apweiler,R. (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**, 741–742.
14. Nugent,T. and Jones,D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.
15. Chatterjee,S. and Mayor,S. (2001) The GPI-anchor and protein sorting. *Cell. Mol. Life. Sci.*, **58**, 1969–1987.
16. Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.
17. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.
18. Briesemeister,S., Rahnenführer,J. and Kohlbacher,O. (2010) Going from where to why–interpretable prediction of protein subcellular localization. *Bioinformatics*, **26**, 1232–1238.
19. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
20. Sharpe,H.J., Stevens,T.J. and Munro,S. (2010) A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell*, **142**, 158–169.
21. Laurila,K. and Vihinen,M. (2011) PROlocalizer: integrated web service for protein subcellular localization prediction. *Amino Acids*, **40**, 975–980.
22. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
23. Wang,G. and Dunbrack,R.L. Jr (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.*, **33**, W94–W98.

Downloaded from http://nar.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

DOI: 10.1093/jnci/djr130
Advance Access publication on April 19, 2011.

© The Author 2011. Published by Oxford University Press. All rights reserved.
For Permissions, please e-mail: journals.permissions@oup.com.

SDHA Loss-of-Function Mutations in *KIT–PDGFRA* Wild-Type Gastrointestinal Stromal Tumors Identified by Massively Parallel Sequencing

Maria A. Pantaleo, Annalisa Astolfi, Valentina Indio, Richard Moore, Nina Thiessen, Michael C. Heinrich, Chiara Gnocchi, Donatella Santini, Fausto Catena, Serena Formica, Pier Luigi Martelli, Rita Casadio, Andrea Pession, Guido Biasco

Manuscript received December 7, 2010; revised March 10, 2011; accepted March 25, 2011.

Correspondence to: Maria A. Pantaleo, Department of Hematology and Oncological Sciences "L.A.Seragnoli," Sant'Orsola-Malpighi Hospital, University of Bologna, Via Massarenti 9, 40138 Bologna, Italy (e-mail: maria.pantaleo@unibo.it).

Approximately 10%–15% of gastrointestinal stromal tumors (GISTs) in adults do not harbor any mutation in the *KIT* or *PDGFRA* genes (ie, KIT/PDGFRA wild-type GISTs). Recently, mutations in *SDHB* and *SDHC* (which encode succinate dehydrogenase subunits B and C, respectively) but not in *SDHA* and *SDHD* (which encode subunits A and D, respectively) were identified in KIT/PDGFRA wild-type GISTs. To search for novel pathogenic mutations, we sequenced the tumor transcriptome of two young adult patients who developed sporadic KIT/PDGFRA wild-type GISTs by using a massively parallel sequencing approach. The only variants identified as disease related by computational analysis were in *SDHA*. One patient carried the homozygous nonsense mutation p.Ser384X, the other patient was a compound heterozygote harboring a p.Arg31X nonsense mutation and a p.Arg589Trp missense mutation. The heterozygous nonsense mutations in both patients were present in germline DNA isolated from peripheral blood. Protein structure analysis indicates that all three mutations lead to functional inactivation of the protein. This is the first report, to our knowle dge, that identifies *SDHA* inactivation as a common oncogenic event in GISTs that lack a mutation in *KIT* and *PDGFRA*.

J Natl Cancer Inst 2011;103:983–987

Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal tumors of the gastrointestinal tract and arise from the interstitial cells of Cajal. In approximately 85% of GISTs, gain-of-function mutations in either the *KIT* gene (which encodes a receptor for stem cell factor) or the platelet-derived growth factor receptor, alpha polypeptide (*PDGFRA*) gene are the oncogenic events that lead to tumor development, resulting in the constitutive ligand-independent activation of the receptor tyrosine kinases encoded by these genes and their downstream signaling pathways (1). Approximately 10% of GISTs in adult patients do not harbor a mutation in either gene (defined as KIT/PDGFRA wild-type). Notably, approximately 85% of GISTs that arise in children are KIT/PDGFRA wild-type and are often associated with a cancer syndrome (2). KIT/PDGFRA wild-type GISTs are often localized to the stomach, are multicentric in origin, and can have an indolent clinical course (3,4). The introduction of inhibitors of the KIT and PDGFRA tyrosine kinases to the therapeutic armamentarium has dramatically changed the medical treatment of GIST patients; as has been widely demonstrated, treatment response with these inhibitors strictly depends on the mutation status of *KIT* and *PDGFRA* (5–10). For example, in the metastatic setting, KIT/PDGFRA wild-type GISTs are more resistant to imatinib and more sensitive to sunitinib than GISTs with a mutant kinase (eg, GISTs with a *KIT* exon 11 mutation) (7–10). KIT/PDGFRA wild-type GISTs also differ from KIT/PDGFRA mutant GISTs in their clinical behavior and underlying genomic background and thus represent a distinct molecular subtype of GIST. Gene expression and gene copy number profiles of KIT/PDGFRA wild-type GISTs differ from those of mutant GISTs (11–15). For example, among GISTs that arise in children and young adults, insulin-like growth factor 1 receptor overexpression is commonly observed in those that are KIT/PDGFRA wild-type but not in those with either mutant kinase (11–13). Moreover, KIT/PDGFRA wild-type GISTs in children and young adults have minimal genomic copy number changes compared with kinase-mutant GISTs, which frequently have gross chromosomal copy number changes, including complete or partial deletion of chromosome arm 14 and deletion of chromosome arms 1p and 22q (14,15). Except for rare case reports of activating mutations in *BRAF* (16) and a recent report of mutations in *SDHB* and *SDHC* (which encode subunits B and C, respectively, of succinate dehydrogenase [SDH]) (17), to our knowledge, no pathogenic mutations have been identified in nonsyndromic GISTs that are *KIT* and *PDGFRA* mutation negative. We searched for novel pathogenic mutations by performing whole-transcriptome next-generation sequencing of sporadic KIT/PDGFRA wild-type GISTs that arose in two young adult patients (GIST_07 and GIST_10). Next generation RNA sequencing is the only approach that allows the complete and thorough identification of all the possible genetic alterations (point mutations, insertions, deletions, and rearrangements) for all genes expressed in a pathological sample. This study was approved by the local institutional ethical committee of Azienda Ospedaliero-Universitaria Policlinico S. Orsola-Malpighi (approval number 113/2008/U/Tess). All patients provided written informed consent.

Poly(A) RNA was isolated from each tumor and subjected to whole-transcriptome paired-end sequencing with the use of a Genome Analyzer IIx system

Downloaded from http://jnci.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

CONTEXT AND CAVEATS

Prior knowledge

Approximately 10% of gastrointestinal stromal tumors (GISTs) in adults and 85% of GISTs in children do not carry a gain-of-function mutation in the receptor tyrosine kinase–encoding *KIT* or *PDGFRA* genes. KIT/PDGFRA wild-type GISTs differ from KIT/PDGFRA mutant GISTs in their response to treatment with kinase inhibitors, clinical behavior, and underlying genomic background.

Study design

Whole-transcriptome next-generation sequencing was used to search for novel pathogenic mutations in sporadic KIT/PDGFRA wild-type GISTs that arose in two young adult patients. Computational analysis was used to determine the likelihood that a mutation is disease related or not.

Contribution

The only variants identified as disease related by computational analysis were in *SDHA*, the gene encoding succinate dehydrogenase subunit A. One patient carried a homozygous nonsense mutation, the other patient was a compound heterozygote harboring a nonsense mutation and a missense mutation. The heterozygous nonsense mutations in both patients were present in germline DNA isolated from peripheral blood.

Implications

SDHA inactivation may be a common oncogenic event in GISTs that lack a mutation in *KIT* and *PDGFRA*.

Limitations

Only two KIT/PDGFRA wild-type GISTs were sequenced.

From the Editors

(Illumina, San Diego, CA) as described in detail in Supplementary Methods (available online), yielding an average of 202 779 378 reads that aligned onto the human reference genome for both patients. Alignments were processed according to a routine technical procedure for calling single-nucleotide variants, which were filtered to exclude known polymorphisms that are annotated in the National Center for Biotechnology Information dbSNP (http://www.ncbi.nlm.nih.gov/snp) and the 1000 Genomes databases, leaving 173 617 and 183 159 putative novel variants for GIST_07 and GIST_10, respectively (Supplementary Table 1, available online). The two patients had nonsynonymous mutations in the coding sequences of the same 261 genes, for a total of 582 novel variants, which we sorted according to a threshold confidence value (Supplementary Methods, available online). This procedure identified nine candidate genes that were mutated in the tumors of both patients (Supplementary Table 2, available online). The mutations in these nine genes were analyzed with SNPs&GO (18), a method for computing the likelihood of a mutation being disease related or not depending on the protein sequence and its functional annotation. This method predicted that only the three mutations in the coding sequence of *SDHA*, which encodes subunit A of SDH, are disease related with a high reliability index (Supplementary Table 2, available online). Massively parallel sequencing analysis revealed that GIST_07 had a C to G transversion at nucleotide 1151 in exon 9 of *SDHA*, a nonsense mutation resulting in the replacement of serine with a stop codon at residue 384 of SDHA, which causes truncation of the peptide chain at residue 383 (p.Ser384X). GIST_10 had two mutations in *SDHA*: 1) a C to T transition at nucleotide 91 in exon 2, a nonsense mutation resulting in the replacement of arginine with a stop codon at residue 31 (p.Arg31X) and 2) a C to T transition at nucleotide 1765 in exon 13, a missense mutation resulting in the replacement of arginine at residue 589 with tryptophan (p.Arg589Trp; Table 1).

SDH (also known as complex II) consists of four subunits: SDHA, SDHB, SDHC, and SDHD (19). Mutant SDH results in dysfunction of complex II of the electron transport chain in mitochondria and, consequently, defective oxidative phosphorylation, which mediates a pseudohypoxic response (ie, the abnormal stabilization of hypoxia-inducible factors [HIFs] under normoxic conditions). Patients with the Carney–Stratakis syndrome, who are predisposed to developing paragangliomas and GISTs, have germline mutations in *SDHB*, *SDHC*, and *SDHD* (20–22). Although SDHA forms a complex with SDHB, SDHC, and SDHD, to our knowledge, no mutations in *SDHA* have been reported in patients with the Carney–Stratakis syndrome or in patients who develop sporadic KIT/PDGFRA wild-type GISTs.

To validate our results and discriminate whether the detected *SDHA* mutations were present in the germline or somatic, we performed targeted exon sequencing of DNA isolated from tumor and peripheral blood of both patients (Supplementary Methods, available online). Patient GIST_07 carried c.1151C>G as a heterozygous germline mutation in blood and as a homozygous mutation in the tumor (Figure 1, A). Single-nucleotide polymorphism array analysis (15) and quantitative polymerase chain reaction analysis of tumor and peripheral blood DNA revealed that there was no statistically significant difference in copy number at the *SDHA* locus between the tumor DNA and the matched peripheral blood DNA, which suggests that the mutation in the tumor was present in homozygosis (Supplementary Figure 1, available online). Quantitative polymerase chain reaction analysis of cDNA synthesized from tumor RNA revealed that the homozygous nonsense mutation in the GIST_07 tumor was associated with a

Table 1. Succinate dehydrogenase subunit A (*SDHA*) mutations

Gene ID	Gene name	Uniprot ID	Genomic coordinate*	Exon	Patient	Allele mutation	Residue mutation†
SDHA	Succinate dehydrogenase (ubiquinone) flavoprotein subunit A	P31040	chr5:288345	9	GIST_07	C>G	S384X
			chr5:276624	2	GIST_10	C>T	R31X
			chr5:304554	13	GIST_10	C>T	R589W

* As in National Center for Biotechnology Information v36.1.

† Residue substitutions are shown with their position in the protein chain.

Downloaded from http://jnci.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

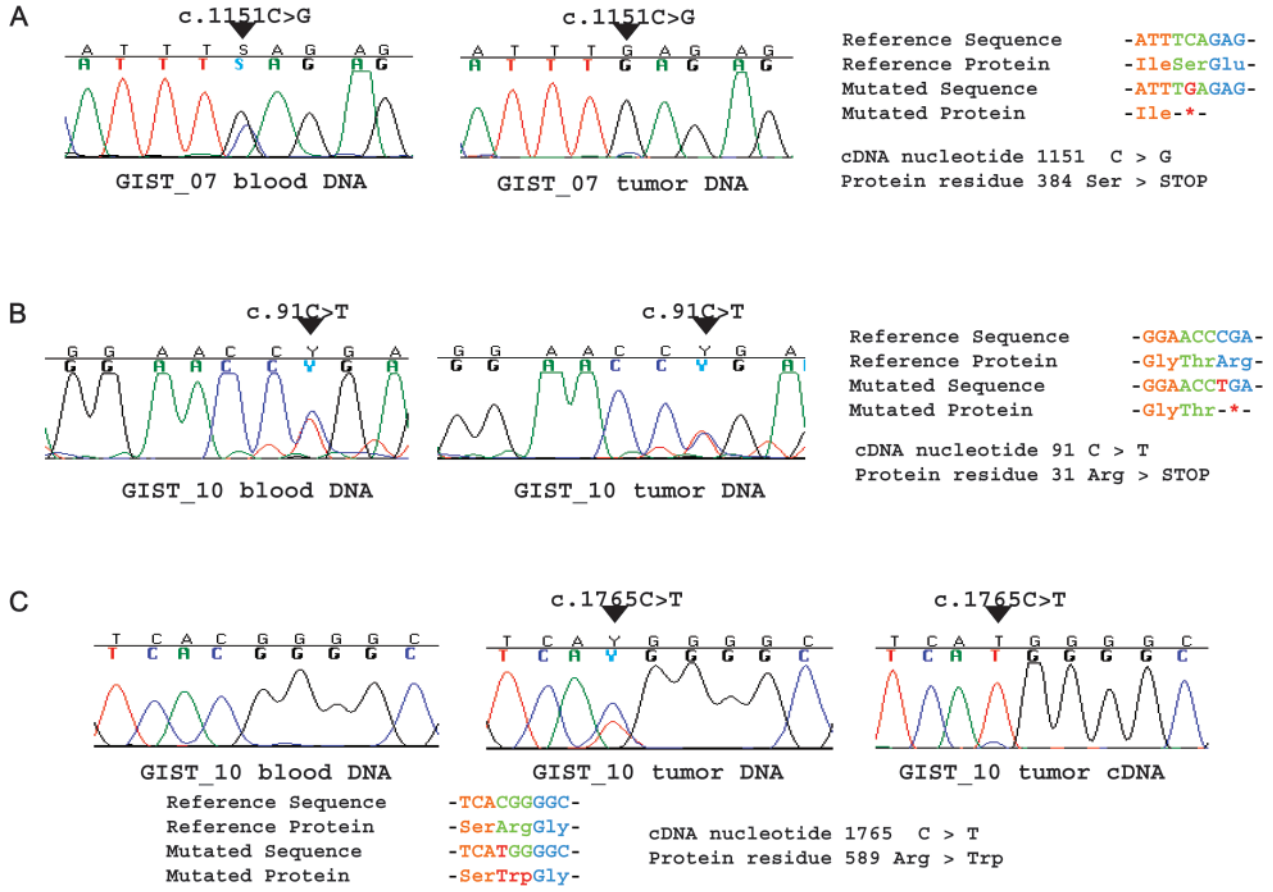


Figure 1. Sequence chromatograms of DNA isolated from two patients with sporadic KIT/PDGFRA wild-type gastrointestinal stromal tumors. **A**) Region harboring the c.1151C>G mutation (p.Ser384X) in patient GIST_07. *Left,* heterozygous mutation in blood DNA; *Right,* homozygous mutation in tumor DNA. **B**) Region harboring the c.91C>T mutation (p.Arg31X) in patient GIST_10. *Left,* heterozygous mutation in blood DNA; *Right,* heterozygous mutation in tumor DNA. **C**) Region harboring the c.1765C>T mutation (p.Arg589Trp) carried by GIST_10 patient. *Left,* wild-type sequence in GIST_10 blood DNA; *Middle,* heterozygous mutation in GIST_10 tumor DNA; *Right,* predominance of expression of the mutated allele in cDNA synthesized from GIST_10 tumor RNA.

marked reduction in the level of SDHA mRNA compared with that in the 14 of 17 KIT/PDGFRA mutant GIST samples for which there was sufficient RNA available for analysis (mean normalized SDHA expression, GIST_07 vs 14 mutant GISTs = 0.89 vs 6.08; fold difference = 6.8, 95% CI = 4.5 to 12.2; $P < .001$, Student t test, two-tailed) (Supplementary Figure 2, available online). A KIT/PDGFRA wild-type GIST from a pediatric patient (GIST_24) had essentially the same fold reduction in SDHA mRNA compared with KIT/PDGFRA mutant GISTs as did GIST_07 (Supplementary Figure 2, available online).

Patient GIST_10 carried c.91C>T as a heterozygous nonsense mutation in both blood and tumor (Figure 1, B), indicating that this patient had a germline genetic alteration of the *SDHA* gene. A second hit that affected *SDHA* in the tumor of this patient was compound heterozygosity for an independent somatic mutation (c.1765C>T), resulting in the Arg589Trp mutation in the mature protein (Figure 1, C). Sequence analysis of cDNA synthesized from tumor RNA revealed that the mutant allele was predominantly expressed in the tumor (Figure 1, C). To understand the effect of the Arg589Trp mutation in the mature SDHA, we computed a three-dimensional model of the mutated subunit by adopting as a template the structure of its porcine counterpart, which is known at atomic resolution (Supplementary Figure 3, available online). Protein structure analysis highlights that, in the wild-type protein, Arg 589 is located in the flavin adenine dinucleotide-binding domain, which is critical for SDHA function. The Arg589Trp mutation results in a side-chain substitution that promotes misfolding of this domain (and, as a consequence, functional inactivation of SDHA) by destabilizing the local polar environment of the wild-type Arg 589 (Supplementary Figure 4, available online).

Together, these results indicate that patients GIST_07 and GIST_10 each carried a first-hit germline mutation in *SDHA*, represented by two different single-base changes, which introduced a stop codon that resulted in a truncated mature protein. A SDHA heterozygous nonsense mutation in the germline of both patients may indicate a neoplastic syndrome that includes KIT/PDGFRA mutation–negative GISTs and potentially other cancers. It would be interesting to know the cancer history of long-term survivors of GISTs that are *PDGFRA* and *KIT* mutation negative.

Recently, Janeway et al. (17) found germline mutations in *SDHB*, *SDHC*, or *SDHD* in six of 38 KIT/PDGFRA wild-type GISTs from pediatric patients with no family history of paraganglioma and, moreover, the loss of SDHB protein expression and complex II activity in KIT/PDGFRA wild-type GISTs with no *SDHB*, *SDHC*, or *SDHD* mutations or deletions from 13

Downloaded from http://jnci.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

pediatric patients. These findings support our view that loss of SDH function plays a role in the pathogenesis of KIT/PDGFRA wild-type GISTs and together with our findings suggest that children or young adults with KIT/PDGFRA wild-type GISTs should be screened for germline or de novo mutations in all four subunits of SDH complex. Disruption of the SDH complex leads to increased expression of HIF-1 alpha and may cause GIST or paraganglioma through similar molecular pathways as seen in renal cell cancers that display loss of von Hippel–Lindau tumor suppressor function (23).

The exact role of mutant SDHA in tumor initiation is poorly understood. Recently, Burnichon et al. (24) identified a germline *SDHA* mutation resulting in p. Arg589Trp that was associated with loss of heterozygosity in a catecholamine-secreting abdominal paraganglioma and suggested that *SDHA* may work as a tumor suppressor gene. The authors showed that this mutant SDHA was associated with loss of enzymatic activity of the SDH complex in tumor tissue and in a yeast model, and, like mutations in *SDHB*, *SDHC*, and *SDHD*, resulted in pseudohypoxia and increased angiogenesis and cell proliferation in vitro. In this study, we found that *SDHA* gene mutations are present in PDGFRA/KIT wild-type GISTs, supporting the hypothesis that SDHA may act as a tumor suppressor in these tumors.

A limitation of this study is the small sample size. A larger number of KIT/PDGFRA wild-type GISTs needs to be evaluated for *SDHA* mutations to reach definitive conclusions about the role of this gene in the development of sporadic KIT/PDGFRA wild-type GISTs. To our knowledge, this is the first report describing germline and somatic loss-of-function mutations in *SDHA* that are linked to the development of sporadic KIT/PDGFRA wild-type GISTs. The finding that inactivation of the SDH complex seems to be an event shared by sporadic and syndromic GISTs that lack mutations in *PDGFRA* and *KIT* may open new avenues for pharmacologic treatments.

References

1. Hirota S, Isozaki K, Moriyama Y, et al. Gain of function mutations of c-kit in human gastrointestinal stromal tumors. *Science.* 1998; 279(5350):577–580.
2. Stratakis CA, Carney JA. The triad of paragangliomas, gastric stromal tumors and pulmonary chondromas (Carney triad), and the dyad of paragangliomas and gastric stromal sarcomas (Carney-Stratakis syndrome): molecular genetics and clinical implications. *J Intern Med.* 2009;266(1):43–52.
3. Janeway KA, Pappo AS. Pediatric gastrointestinal stromal tumor. *Hematol Oncol Clin North Am.* 2009;23(1):15–34.
4. Miettinen M, Lasota J, Sobin LH. Gastrointestinal stromal tumors of the stomach in children and young adults: a clinicopathologic, immunohistochemical, and molecular genetic study of 44 cases with long-term follow-up and review of the literature. *Am J Surg Pathol.* 2005;29(10):1373–1381.
5. Demetri GD, Von Mehren M, Blanke CD, et al. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *N Engl J Med.* 2002;347(7):472–480.
6. Demetri GD, van Oosterom AT, Garrett CR, et al. Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of imatinib a randomised controlled trial. *Lancet.* 2006;368(9544):1329–1338.
7. Heinrich MC, Corless CL, Demetri GD, et al. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumors. *J Clin Oncol.* 2003;21(23): 4342–4349.
8. Debiec-Rychter M, Dumez H, Judson I, et al. Use of c-KIT/PDGFRA mutational analysis to predict the clinical response to imatinib in patients with advanced gastrointestinal stromal tumours entered on phase I and II studies of the EORTC Soft Tissue and Bone Sarcoma Group. *Eur J Cancer.* 2004;40(5):689–695.
9. Heinrich MC, Corless CL, Blanke CD, et al. Molecular correlates of imatinib resistance in gastrointestinal stromal tumors. *J Clin Oncol.* 2006;24(29):4764–4774.
10. Heinrich MC, Maki RG, Corless CL, et al. Primary and secondary kinase genotypes correlate with the biological and clinical activity of sunitinib in imatinib-resistant gastrointestinal stromal tumor. *J Clin Oncol.* 2008;26(33): 5352–5359.
11. Pantaleo MA, Astolfi A, Di Battista M, et al. Insulin-like growth factor 1 receptor (IGF1r) expression in wild-type GIST: a potential novel therapeutic target. *Int J Cancer.* 2009; 125(12):2991–2994.
12. Tarn C, Rink L, Merkel E, et al. Insulin-like growth factor 1 receptor is a potential therapeutic target for gastrointestinal stromal tumors. *Proc Natl Acad Sci U S A.* 2008; 105(24):8387–8392.
13. Janeway KA, Zhu MJ, Barretina J, Perez-Atayde A, Demetri GD, Fletcher JA. Strong expression of IGF1R in pediatric gastrointestinal stromal tumors without IGF1R genomic amplification. *Int J Cancer.* 2010; 127(11):2718–2722.
14. Agaram NP, Laquaglia MP, Ustun B, et al. Molecular characterization of pediatric gastrointestinal stromal tumors. *Clin Cancer Res.* 2008;14(10):3204–3215.
15. Astolfi A, Nannini M, Pantaleo MA, et al. A molecular portrait of gastrointestinal stromal tumors: an integrative analysis of gene expression profiling and high-resolution genomic copy number. *Lab Invest.* 2010;90(9):1285–1294.
16. Agaram NP, Wong GC, Guo T, et al. Novel V600E BRAF mutations in imatinib-naive and imatinib-resistant gastrointestinal stromal tumors. *Genes Chromosomes Cancer.* 2008; 47(10):853–859.
17. Janeway KA, Kim SY, Lodish M, et al. Defects in succinate dehydrogenase in gastrointestinal stromal tumors lacking KIT and PDGFRA mutations. *Proc Natl Acad Sci U S A.* 2011; 108(1):314–318.
18. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat.* 2009;30(8):1237–1244.
19. Cecchini G, Schröder I, Gunsalus RP, Maklashina E. Succinate dehydrogenase and fumarate reductase from Escherichia coli. *Biochim Biophys Acta.* 2002;1553(1–2):140–157.
20. McWhinney SR, Pasini B, Stratakis CA; International Carney Triad and Carney-Stratakis Syndrome Consortium. Familial gastrointestinal stromal tumors and germ-line mutations. *N Engl J Med.* 2007;357(10): 1054–1056.
21. Pasini B, McWhinney SR, Bei T, et al. Clinical and molecular genetics of patients with the Carney-Stratakis syndrome and germline mutations of the genes coding for the succinate dehydrogenase subunits SDHB, SDHC, and SDHD. *Eur J Hum Genet.* 2008;16(1):79–88.
22. Pasini B, Stratakis CA. SDH mutations in tumorigenesis and inherited endocrine tumours: lesson from the phaeochromocytoma-paraganglioma syndromes. *J Intern Med.* 2009; 266(1):19–42.
23. Linehan WM, Srinivasan R, Schmidt LS. The genetic basis of kidney cancer: a metabolic disease. *Nat Rev Urol.* 2010;7(5):277–285.
24. Burnichon N, Brière JJ, Libé R, et al. SDHA is a tumor suppressor gene causing paraganglioma. *Hum Mol Genet.* 2010;19(15): 3011–3020.

Funding

Novartis Oncology, Origgio, Italy; VA Merit Review grant (to M.C.H.); GIST Cancer Research Fund (to M.C.H.); Life Raft Group (to M.C.H.); Novartis Oncology (to M.C.H.).

Notes

MAP and AA contributed equally to the work. MCH is a consultant for, and has equity interest in, Molecular MD for *KIT* and *PDGFRA* genotype analysis in GIST.

Special thanks to Manuela Iero (Novartis Oncology, Origgio Italy); Monica Di Battista, Valerio Di Scioscio, Paola Paterini, Claudio Ceccarelli, Paolo Castellucci, Michelangelo Fiorentino, Pietro Fusaroli, Maristella Saponara,

Downloaded from http://jnci.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014

Cristian Lolli, Margherita Nannini, Alessandra Maleddu, Anna Mandrioli, Maria Caterina Pallotti, Franco W. Grigioni, Antonio D. Pinna, Ilaria Caldarera and the staff of Gist Study Group (Gist Study Group Bologna, Italy); Diane Miller, Yongjun Zhao, Thomas Zeng, Tina Wong, Richard Varhol and the staff of Genome Sciences Centre (Genome Sciences Centre, Vancouver, Canada). The sponsor did not have role in the design of the study, but approved it before providing financial support. The sponsor approved the submitted version of the article and the revised version for publication. The sponsor did not have role in the collection, analysis, or interpretation of the data or the writing of the article.

Affiliations of authors: Department of Hematology and Oncology Sciences "L&A Seràgnoli" (MAP, GB), Pathology Unit (DS), and Department of Surgery and Transplantation (FC), Sant'Orsola-Malpighi Hospital, University of Bologna, Bologna, Italy; "Giorgio Prodi" Cancer Research Center, University of Bologna, Bologna, Italy (MAP, AA, VI, SF, AP, GB); Biocomputing Group, Department of Biology, University of Bologna, Bologna, Italy (VI, PLM, RC); Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada (RM, NT); Department of Medicine, Portland VA Medical Center and Oregon Health and Science University Knight Cancer Institute, Portland, OR (MCH); Novartis Oncology, Origgio, Italy (CG); Pediatric Unit, Department of Gynecological, Obstetric and Pediatric Sciences, University of Bologna, Bologna, Italy (SF, AP).

Downloaded from http://jnci.oxfordjournals.org/ at Università di Bologna - Sistema Bibliotecario d'Ateneo on March 25, 2014