**SIFT**

SIFT è uno score basato su omologia di sequenze che ordina le sostituzioni aminoacidiche da tollerate ad intollerate e predice se una sostituzione aminoacidica in una proteina può avere un effetto fenotipico. SIFT è basato sulla premessa che l’evoluzione di una proteina è correlata con la funzione della proteina.

COME FUNZIONA?

SIFT prende una sequenza ed utilizza informazioni dovute da multipli allineamenti per predire le sostituzioni tollerate e deleterie per ogni posizione sulla sequenza query. Gli step sono:

1) cerca sequenze simili

2) sceglie le sequenze in maggior relazione che possono avere funzioni simili a quelle espresse nella sequenza query

3) effettua gli allineamenti delle sequenze scelte

4) calcola le probabilità normalizzate per tutte le possibili sostituzioni dall’allineamento.

VALUTAZIONE SCORE:

Posizioni con prob. Normalizzata minore di 0.05 sono predette deleterie. Se maggiori o uguali a 0.05 sono predette benigne.

SIFT < 0.05 = BENIGN

SIFT >= 0.05 = DELETERIOUS

ALTRI SCORE:

**SIFT\_converted\_rankscore**: SIFTori scores were first converted to SIFTnew=1-SIFTori, then ranked among all SIFTnew scores in dbNSFP. The rankscore is the ratio of the rank the SIFTnew score over the total number of SIFTnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The rankscores range from 0.00963 to 0.91219.

**POLYPHEN-2**

PolyPhen-2 è uno score usato per predire tutti i possibili impatti di una sostituzione aminoacidica sulla struttura e la funzione di una proteina. Questa predizione è basata su un numero di features che comprendono la sequenza, la filogenetica e le informazioni strutturali che caratterizzano la sostituzione. Data una sostituzione amino acidica, polyphen-2 estra le features di cui prima e le da in pasto ad un classificatore.

A substitution may occur at a specific site, e.g., active or binding, or in a non-globular, e.g., trans-membrane, region. PolyPhen-2 tries to identify a query protein as an entry in the human proteins subset of UniProtKB/Swiss-Prot database and use the feature table (FT) section of the corresponding entry.

COME FUNZIONA?

PolyPhen-2 BLASTs query sequence against protein structure database (PDB) and by default retains all hits that meet the given criteria:

* sequence identity threshold is set to 50%, since this value guarantees the conservation of basic structural characteristics
* minimal hit length is set to 100
* maximal number of gaps is set to 20

By default, a hit is rejected if its amino acid at the corresponding position differs from the amino acid in the input sequence. The position of the substitution is then mapped onto the corresponding positions in all retained hits. Hits are sorted according to the sequence identity or E-value of the sequence alignment with the query protein.

Further analysis performed by PolyPhen-2 is based on the use of several structural parameters. Importantly, although all parameters are reported in the output, only some of them are used in the final decision rules.

PolyPhen-2 uses DSSP (Dictionary of Secondary Structure in Proteins) database to get the following structural parameters for the mapped amino acid residues:

* Secondary structure (according to the DSSP nomenclature)
* Solvent accessible surface area (absolute value in Å²)
* Phi-psi dihedral angles

PREDIZIONE:

Metodo di classificazione supervisionata tramite classificatore Naive bayes. Sono stati usati due db per il train ed il test di PolyPhen2:

* **HumDiv**: was compiled from all damaging alleles with known effects on the molecular function causing human Mendelian diseases, present in the UniProtKB database, together with differences between human proteins and their closely related mammalian homologs, assumed to be non-damaging.
* **HumVar**: consisted of all human disease-causing mutations from UniProtKB, together with common human nsSNPs (MAF>1%) without annotated involvement in disease, which were treated as non-damaging.

Quindi la predizione di PolyPhen è basata su HumanVar ed identifica:

D = Probably Damaging (Hvar\_score [0.909 -1] o rankscore [0.62797 – 0.97092]

P = Possibly Damaging (Hvar [0.447 – 0.908] o rankscore [0.44195 – 0.62727]

B = Benign (Hvar [0 – 0.446] o rankscore [0.01257 – 0.44151])

Il cutoof per la classificazione binaria è 0.5 per HVAR o 0.45833 per il rankscore:

NEUTRAL = HVAR <= 0.5

DAMAGING = HVAR > 0.5

ALTRI SCORE:

**Polyphen2\_HDIV\_score**: Polyphen2 score based on HumDiv, i.e. hdiv\_prob. The score ranges from 0 to 1. Multiple entries separated by ";", corresponding to Uniprot\_acc\_Polyphen2.

**Polyphen2\_HDIV\_rankscore**: Polyphen2 HDIV scores were first ranked among all HDIV scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.02634 to 0.89865.

**Polyphen2\_HVAR\_score**: Polyphen2 score based on HumVar, i.e. hvar\_prob. The score ranges from 0 to 1. Multiple entries separated by ";", corresponding to Uniprot\_acc\_Polyphen2.

**Polyphen2\_HVAR\_rankscore**: Polyphen2 HVAR scores were first ranked among all HVAR scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.01257 to 0.97092.

**LRT SCORE: Likelihood Ratio Test**

Si basa sul concetto di conservazione evolutiva. In genere, le mutazioni deleterie o damaging dovrebbero essere eliminate da un processo di selezione naturale per assicurare sopravvivenza a lungo termine della specie. Tuttavia, non sempre è così e nell’uomo sono presenti molte mutazioni estremamente dannose e deleterie, quindi disease-causing.

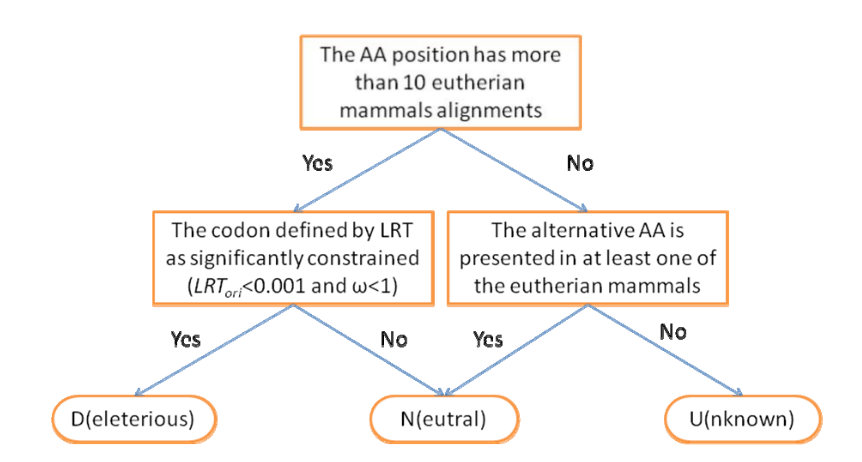
La necessità è quindi quella di sviluppare modelli evolutivi in grado di prevedere mutazioni deleterie per l’uomo. Le mutazioni che alterano le sequenze conservate nell’evoluzione sono probabilmente deleterie e hanno un impatto negativo sulla fitness. Per distinguere posizioni aminoacidiche constrained (vincolanti, VEDI GLOSSARIO a fondo pagine) enon constrained (vincolanti), è stato usato per questo score un database composto da 32 specie vertebrate. Per circa 19000 geni umani sono stati generati multipli allineamenti di sequenze proteiche codificanti ortologhe. Valutando queste sequenze, si è visto che la frequenza di sostituzioni sinonime è di circa 12.2 attraverso tutte le specie e 4.78 tra gli euteria clade, i mammiferi placentati, similmente a precedenti studi. Come si può capire da questi dati, la divergenza tra le specie implica che le posizioni aminoacidiche non vincolanti dovrebbero essere raramente conservate tra le specie mentre le posizioni che alterano aminoacidi conservati sono molto probabilmente deleterie.

COME FUNZIONA:

LRT confronta la la probabilità del dato sotto un modello conservativo, consentendo qualsiasi livello di vincolo selettivo, relativo ad un modello neutrale, dove non vi sono differenze tra frequenze di sostituzione sinonime e non sinonime. Quindi il modello LRT confronta il modello nullo in cui ogni codone sta evolvendo neutralmente (con nessuna differenza tra freq. Di sostituzioni sinonime e non sinonime) rispetto al modello alternativo in cui il codone è evoluto sotto una selezione negativa regolata da un parametro libero per il rapporto (freq. Non sinonime)/(freq. Sinonime).  
Lo score LRT è un p-value della Likelihood Ratio test di codoni vincolati. Ogni score è associato ad un rapporto OMEGA --> (freq. Non sinonime)/(freq. Sinonime) ed un allineamento aminoacidico di 31 specie per il codone di test. Se il codone è più vincolato rispetto al modello neutrale allora OMEGA<1 altrimenti OMEGA>=1.

PREDIZIONE:

Lo score varia tra 0 e 1 e score più grandi stanno a significare che il codone è più conservato o che una Non Sinonima sia più probabilmente deleteria. Lo score si ottiene come da figura:



In conclusione:

* D = DAMAGING: se soddisfa tre criteri: 1) LRT definisce il codone in cui cade la NS (non sinonima) come significativamente conservato (OMEGA < 1). 2) il sito deve avere più di 10 allineamenti di mammiferi euteri. 3) l’aminoacido alternativo non è presente in alcun mammifero euterio.
* N = NEUTRAL: due criteri da soddisfare: 1) l’aminoacido alternativo è presente in almeno uno dei mammiferi euteri o 2) da un codone definito da LRT come non significativamente conservato (OMEGA >= 1) e con almeno 10 allineamenti di mammiferi euterii.
* U = UNKNOWN: altrimenti definito come unknown.

ALTRI SCORE:

**LRT\_converted\_rankscore**: LRTori scores were first converted as LRTnew=1-LRTori\*0.5 if Omega<1, or LRTnew=LRTori\*0.5 if Omega>=1. Then LRTnew scores were ranked among all LRTnew scores in dbNSFP. The rankscore is the ratio of the rank over the total number of the scores in dbNSFP. The scores range from 0.00162 to 0.84324.

**LRT\_Omega**: estimated nonsynonymous-to-synonymous-rate ratio (Omega, reported by LRT)

**MUTATIONTASTER**

MutTast valuta le varianti in base al loro potenziale disease-causing. Il tool effettua una serie di test in silico per stimare l’impatto della variante sul gene/prodotto. I test sono effettuati sia a livello della pèroteina che del DNA, per questo motivo non è limitato solo alle missenso ma anche alle sinonime e le introniche.

Il software prima scarta le mutazioni note ed i polimorfismi tramite confronto con i database integrati. Le rimanenti SNPs vengono testate in accordo all’alterazione genica che essi causano:

* Sinonime silenti o alterazioni introniche che non portano a cambi aminoacidici.
* Mutazioni che riguardano il singolo aminoacido.
* Mutazioni che causano complessi cambi nella sequenza aminoacidica (come indels).

PREDIZIONE:

Lo score varia tra 0 e 1.

A = Disease causing automatic

D = Disease causing

N = polymorphism

P = polymorphism automatic

Il cutoff tra D ed N è pari a 0.5 per Mtnew e 0.31713 per il rankscore.

ALTRI SCORE:

**MutationTaster\_converted\_rankscore:** The MTori scores were first converted: if the prediction is "A" or "D" MTnew=MTori; if the prediction is "N" or "P", MTnew=1-MTori. Then MTnew scores were ranked among all MTnew scores in dbNSFP. If there are multiple scores of a SNV, only the largest MTnew was used in ranking. The rankscore is the ratio of the rank of the score over the total number of MTnew scores in dbNSFP. The scores range from 0.08979 to 0.81033.

**MutationTaster\_model**: MutationTaster prediction models.

**MutationTaster\_AAE**: MutationTaster predicted amino acid change.

**MUTATIONASSESSOR**

Valuta diverse caratteristiche che inficiano sulla struttura della proteina e quindi sulla sua funzionalità prendendo in considerazione anche il livello biomolecolare.

Si parte dall’assunzione per cui molte mutazioni sono state provate in evoluzione in ogni posizione della sequenza sufficientemente spesso così che si possano osservare le distribuzioni dei residui in posizioni allineate di sequenze omologhe che riflettano il vincolo funzionale di questi residui. Questo significa che i residui evolutivi sfavorevoli sono non osservati o osservati meno frequentemente rispetto ai residui neutrali o a quelli critici, mentre i residui criticamente importanti sono conservati in diversi profili di evoluzione (parologhi o ortologhi).

Queste osservazioni compongono la base per convertire le frequenze osservate in una stima numerica dell’impatto funzionale di una mutazione. Si usa l’entropia per misurare la conservazione dei residui e stimare l’impatto della mutazione valutando le differenze di entropia causate dalla mutazione. Quindi la frequenza dei residui nell’allineamento determina l’entropia ed il risultato dello score.

PREDIZIONE:

H o M = High

L o N = Non Functional

I cutoffs per lo score Maori sono:

H - M = 3.5

M - L = 1.935

L - N = 0.8

Mentre per il RankScore sono rispettivamente 0.92922 , 0.51944 e 0.19719

ALTRI SCORE:

**MutationAssessor\_UniprotID**: Uniprot ID number provided by MutationAssessor.

**MutationAssessor\_variant**: AA variant as to MutationAssessor\_UniprotID.

**MutationAssessor\_score**: MutationAssessor functional impact combined score (MAori). The score ranges from -5.135 to 6.49 in dbNSFP.

**MutationAssessor\_rankscore**: MAori scores were ranked among all MAori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MAori scores in dbNSFP. The scores range from 0 to 1.

**FATHMM\_SCORE e FATHMM-MKL SCORE:**

Lo score viene ottenuto in questo caso tramite Hidden Markov Model. La procedura per determinare le conseguenze funzionali sulla funzione della proteina è la seguente:

* Vengono cercate le sequenze omologhe.
* Contemporaneamente allo step precedente viene costruito un primo HMM.
* Viene costruito un db contenente un’annotazione sul dominio della proteina.
* Le info sul dominio della proteina vengono estratte solo se significative (si applica un test statistico come e-value).
* Il guadagno viene poi calcolato in corrispondenza degli stati che matchano il modello HMM estratto precedentemente.
* In seguito si interrogano le probabilità aminoacidiche modellate dall’HMM più informativo e si assume che una riduzione nella probabilità dell’aminoacido indichi un potenziale impatto negativo sulla funzione della proteina laddove un guadagno della probabilità indica invece una sostituzione più favorevole.

PREDIZIONE:

Lo score varia tra -16.13 e 10.64. Più piccolo è lo score, più è probabile che lo score sia damaging.

ALTRI SCORES:

**FATHMM\_converted\_rankscore**: FATHMMori scores were first converted to FATHMMnew=1-(FATHMMori+16.13)/26.77, then ranked among all FATHMMnew scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of FATHMMnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

**FATHMM\_pred**:

D = Damaging → FATHMMori score <= -1.5

T = Tolerated → FATHMMori > -1.5

Multiple predictions separated by ";", corresponding to Ensembl\_proteinid

**fathmm-MKL\_coding\_score**:

Questo score è un’evoluzione del precedente score in cui vengono aggiunte le informazioni da ENCODE Project Consortium che include informazioni su trascritti ed RNA anche non codificante, su siti di legame e struttura della cromatina. Lo score varia tra 0 e 1.

PREDIZIONE:

D = Deleterio → fathmm-MKL > 0.5

B = Benigna o Neutrale → fathmm-MKL < 0.5

Scores vicini a 0 o ad 1 sono quelli a più alta confidenza. Gli score sui codificanti sono stati allenati usando 10 gruppi di features.

**fathmm-MKL\_coding\_rankscore**: fathmm-MKL coding scores were ranked among all fathmm-MKL coding scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of fathmm-MKL coding scores in dbNSFP.

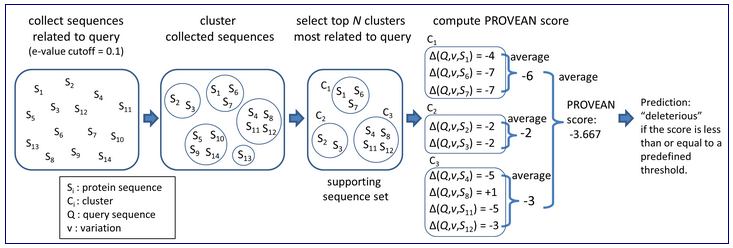
**fathmm-MKL\_coding\_pred**: If a fathmm-MKL\_coding\_score is >0.5 (or rankscore >0.28317) the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "N(EUTRAL)".

**fathmm-MKL\_coding\_group**: the groups of features (labeled A-J) used to obtained the score.

**PROVEAN SCORE:**

Predizione dell’effetto funzionale di una proteina per una data sostituzione. PROVEAN è in grado di fornire uno score per variazioni quali singole o multiple sostituzioni, delezioni ed inserzioni.

Lo score funziona come di seguito:



Data la query, vengono estratte le sequenze che hanno il 75% di identità globale. Le sequenze vengono poi clusterizzate ed infine sono selezionati i primi 30 cluster con punteggio maggiore che supportano la sequenza. Da questi si calcola lo score PROVEAN.

PREDIZIONE:

Lo score varia tra -14 e 14. Più lo score è piccolo, più è damaging. La soglia binaria è settata come di seguito:

D = Damaginng → PROVEAN <= -2.5

N = Neutral → PROVEAN > -2.5

ALTRI SCORES:

**PROVEAN\_converted\_rankscore**: PROVEANori were first converted to PROVEANnew=1-(PROVEANori+14)/28, then ranked among all PROVEANnew scores in dbNSFP. The rankscore is the ratio of the rank the PROVEANnew score over the total number of PROVEANnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

**VEST SCORE:**

Il Variant Effect Scoring Tool è un metodo per prioritizzare varianti missenso che alterano l’attività della proteina. VEST usa un algoritmo di machine learning supervisionato tramite Random Forest per identificare queste mutazioni. Il training set è stato preso selezionando una classe positiva di varianti missenso dal Human Gene Mutation Database e una classe negativa di varianti comuni prese dall’Exome Sequencing Project (ESP). Frequenze alleliche >1% in questo lavoro sono considerate comuni. Tutte le mutazioni sono descritte da un insieme di 86 features.

Sebbene VEST sia designato per identificare mutazioni missenso che alterano l’attività della proteina, la sua costruzione include il contesto addizionale di differenza selettiva tra varianti causative di malattia e varianti comuni.

La classe della malattia ottenuta da HGMD viene arricchita per le varianti sotto selezione di purificazione, laddove la classe neutrale dall ESP include varianti comuni sotto una selezione positiva nella popolazione umana in aggiunta a varianti realmente neutrali. Quindi, VEST può prendere mutazioni e classificarle in base alla loro similarità rispetto a quelle presenti nel modello malattia dell’HGMD e quelle neutrali del modello ESP.

PREDIZIONE:

Lo score varia tra 0 e 1. Più è alto lo score, più la mutazione ha probabilità di essere patogenica.

ALTRI SCORE:

**VEST3\_rankscore**: VEST3 scores were ranked among all VEST3 scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of VEST3 scores in dbNSFP. In case there are multiple scores for the same variant, the largest score (most damaging) is presented. The scores range from 0 to 1.

**MetaSVM SCORE:**

Lo score MetaSVM usa Support vector Machine usando 10 score (SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, GERP++, MutationTaster, MutationAssessor, FATHMM, LRT, Siphy, PhyloP) e la massima frequenza osservata nelle popolazioni 1000 genomi. Valori più grandi indicano che l’SNV è più probabile essere damaging. In dbNSFP lo score varia tra -2 e 3.

PREDIZIONE:

T = Tolerated → MetaSVM < 0

D = Damaging → MetaSVM > 0

Il cutoff per il Rankscore è pari a 0.82268

ALTRI SCORE:

**MetaSVM\_rankscore**: MetaSVM scores were ranked among all MetaSVM scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaSVM scores in dbNSFP. The scores range from 0 to 1.

**MetaLR SCORE:**

Questo score è come il MetaSVM solo che invece di utilizzare support vector machine sfrutta la regressione logistica. Lo score varia tra 0 e 1 e valori più alti indicano che la SNV è più probabilmente patologica

PREDIZIONE:

T = Tolerated → MetaLR < 0.5

D = Damaging → MetaLR > 0.5

Il cutoff per il Rankscore è pari a 0.81113

ALTRI SCORE:

**MetaLR\_rankscore**: MetaLR scores were ranked among all MetaLR scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaLR scores in dbNSFP. The scores range from 0 to 1.

**Reliability\_index**: Number of observed component scores (except the maximum frequency in the 1000 genomes populations) for MetaSVM and MetaLR. Ranges from 1 to 10. As MetaSVM and MetaLR scores are calculated based on imputed data, the less missing component scores, the higher the reliability of the scores and predictions.

**CADD SCORE:**

Combine Adnnotation-Dependent Depletion. Anche questo score usa altri score per identificare un’unica annotazione in merito alla patogenicità di varianti in regioni codificanti e non codificanti sia SNV che piccoli Indel. La base del CADD è quella di mettere in contrasto le annotazioni di alleli derivati fissati o più o meno fissati nell’uomo con quelli di varianti simulate. Le varianti deleterie sono scartate dalla selezione naturale nei modelli fissati ma non in quelli simulati. Quindi CADD fornisce una misura di dannosità correlata sia alla funzionalità molecolare che alla patogenicità. Anche questo include score come i precedenti oltre che scores di conservazione e di funzionalità molecolare.

Lo score varia tra -7.535037 a 35.788538 in dbNSFP. Più lo score è grande, più la variante risulta essere patogenica.

PREDIZIONE:

La predizione di patogenicità può essere vista solo tramite la grandezza dello score, più utile nel caso di rankscore.

ALTRI SCORE:

**CADD\_raw\_rankscore**: CADD raw scores were ranked among all CADD raw scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of CADD raw scores in dbNSFP.

**CADD\_phred**: CADD phred-like score. This is phred-like rank score based on whole genome CADD raw scores. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5 for details. The larger the score the more likely the SNP has damaging effect.

**DANN SCORE:**

Il DANN score (Deleterious Annotation of variants using Neural Network, DANN) è una rete neurale multilayer ovvero in cui i diversi layer vengono posti tra l’input e l’output a scapito però del tempo computazionale. I risultati sono migliori del CADD che invece utilizza un modello lineare come SVM. Inoltre il test è stato effettuato non solo paragonandolo con il CADD ma anche usando un modello di regressione logistica che mostra comunque migliori risultati del CADD.

Lo score varia tra 0 e 1. Un numero prossimo all’1 indica che la variante è patogenica

PREDIZIONE:

La predizione di patogenicità può essere vista solo tramite la grandezza dello score, più utile nel caso di rankscore.

ALTRI SCORES:

**DANN\_rankscore**: DANN scores were ranked among all DANN scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of DANN scores in dbNSFP

**EIGEN SCORE:**

Eigen è un approccio spettrale (?) all’annotazione funzionale di varianti genetiche in regioni codificanti e non codificanti (come quelle disponibili dai progetti ENCODE e Roadmap Epigenomics), e li combina in una singola misura di importanza funzionale.

Sfrutta un approccio non supervisionato e, a differenza della maggior parte dei metodi, non è basato su alcun training dataset. Eigen produce stime di accuratezza predittiva per ogni score di annotazione funzionale e successivamente usa queste stime di accuratezza per ricavare degli score di aggregazione funzionale per le varianti di interesse come una combinazione lineare pesata di annotazioni individuali. I meta-score risultanti hanno una buona capacità discriminante utilizzando le malattie associate e le varianti presunte benigne da studi pubblici (sia per malattie Mendeliane che complesse). Lo score Eigen è particolarmente utile nella prioritizzazione di varianti probabilmente causali in una regione di interesse quando è combinata con dati genetici a livello di popolazione nel framework di un modello gerarchico.

Inoltre, un vantaggio importante dell’Eigen score è che può essere adattato ad uno specifico tessuto o tipo di cellula.

(Come si interpreta lo score? Leggere bene l’articolo)

ALTRI SCORE:

**Eigen-raw**: Eigen score for coding SNVs. A functional prediction score based on conservation, allele frequencies, and deleteriousness prediction using an unsupervised learning method

**Eigen-phred**: Eigen coding score in phred scale.

**Eigen-raw\_rankscore**: Eigen-raw scores were ranked among all Eigen-raw scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of Eigen-raw scores in dbNSFP.

**Eigen-PC-raw**: Eigen PC score for genome-wide SNVs. A functional prediction score based on conservation, allele frequencies, deleteriousness prediction (for missense SNVs) and epigenomic signals (for synonymous and non-coding SNVs) using an unsupervised learning method

**Eigen-PC-raw\_rankscore**: Eigen-PC-raw scores were ranked among all Eigen-PC-raw scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of Eigen-PC-raw scores in dbNSFP.

**GENECANYON SCORE:**

Anche il GeneCanyon score è uno score di predizione funzionale delle regioni. Questo score, così come anche gli altri, sono ottenuti non solo tramite predizioni computazionali come gli score di conservazione, ma anche utilizzando gli esperimenti di high throughput come quelli del progetto ENCODE. GeneCanyon è un metodo di annotazione Whole Genome che effettua un apprendimento statistico non supervisionato usando 22 annotazioni computazionali e annotazioni sperimentali che forniscono una misura del potenziale funzionale di ogni posizione nel genoma umano.

In sostanza, per ogni posizione del genoma calcola la probabilità P(Z = 1|A) dove A indica l’annotazione e sotto l’assunzione che le 22 annotazioni A scelte, siano indipendenti. Scegliendo un cutoff pari 0.5 per lo score, se ne deduce che il 33,3% del genoma umano è predetto essere funzionale e sottoposto ad attività biochimica. Quindi, per ogni posizione, più il valore è vicino ad 1 più la posizione risulta essere funzionale. Per utilizzare questo score è comunque utile inserire un gene intero o una regione intera di analisi (fino a 3 MB).

PREDIZIONE:

FUNCTIONAL = score più vicino ad 1

LESS\_FUNCTIONAL = score più vicino a 0

ALTRI SCORE:

**GenoCanyon\_score\_rankscore:** GenoCanyon\_score scores were ranked among all integrated fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GenoCanyon\_score scores in dbNSFP.

**FITCONS SCORE:**

Il fitCons è un metodo computazionale per stimare la probabilità che una mutazione puntuale ad ogni posizione in un genoma influenzerà la fitness. Questi fitness Consequence (fitCons) score servono come mesure basate sull’evoluzione di potenziali funzioni genomiche.

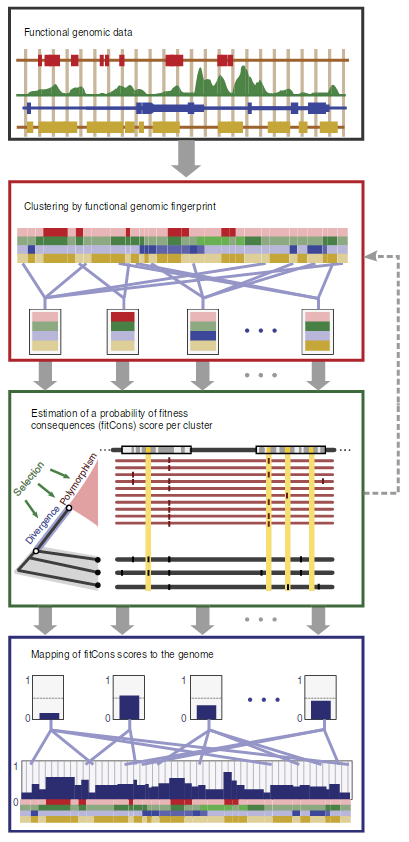
Questo approccio consiste nel clusterizzare le posizione genomiche in gruppi che esibiscono differenti fingerprints sulla base di dati genomici funzionali high-throughput (Mass Array?), per poi stimare la probabilità di fitness consequence per ogni gruppo da patterns associati di polimorfismi genetici e divergenze.

Il metodo statistico usato è chiamato Inference of Natural Selection from Interspersed Genomically Coherent Elements (INSIGHT) il quale mette in contrasto patterns di polimorfismi e la loro divergenza per una collezione di siti genomici sparsi con quelli evolutivi neutrali nelle vicinanze, tenendo in considerazione selezione positiva e negativa. Per questo motivo il metodo integra sia dati funzionali che evolutivi nella caratterizzazione dell’importanza del potenziale funzionale di una regione genomica.

Il metodo segue questa strategia:

* Si inizia con data sets di genome-wide functional ottenuti da ogni tipo cellulare (prima figura).
* Nel secondo step vengono prima clusterizzate le posizioni genomiche a seconda del loro fingerprints genomico funzionale (seconda figura). La suddivisione è concentrata su tre tipi di dati altamente informativi ovvero Dnase I (digestione e sequenziamento), RNA-seq e ChIP-seq che descrivono modificabilità dell’istone. Poi ogni gruppo è stato ulteriormente suddiviso in sotto classi ed infine annotate per verificare che cadessero in sequenze codificanti per proteine. In totale si ottengono 624 classi genomiche funzionali distinte. Poi questo clustering è stato applicato sulle tre linee cellulari.
* Nel terzo step si usa INSIGHT per stimare le probabilità di fitness mutazionale.
* Infine, nel quarto step assegniamo ad ogni posizione nucleotidica nel genoma lo score stimato per la corrispondente classe genomica funzionale.

Di seguito le immagini per ogni step:



FitCons è stato generato su tre linee cellulari umane sulla base di dati pubblici in ENCODE. In confronto agli scores tradizionali risulta più potente per i regolatori *cis*. In aggiunta, questo score mostra che il 4.2-7.5 % dei nucleotidi nel genoma umano hanno una fitness influenzata dalla divergenza chimpazee-umano e suggeriscono che un recente turnover evoluzionario ha avuto un impatto limitato sul contenuto funzionale del genoma. Quindi fitCons predice la frazione di posizioni genomiche appartenenti a specifiche classi funzionali (definite dall’epigenomica come **fingerprint**) che sono sotto pressione selettiva.

Lo score è stato calcolato su queste tre linee cellulari:

1. Cellule endoteliali del cordone ombellicale (HUVEC).
2. Cellule staminali embrionali umane H1 (H1 hESCs).
3. Cellule linfoblastoidi (GM12878).

PREDIZIONE:

Lo score varia tra 0 e 1, con score maggiori che indicano che una più alta proporzione di siti nucleici della classe funzionale a cui appartiene la posizione genomica sono sotto pressione selettiva, e quindi molto più probabile siano importanti da un punto di vista funzionale.

FUNCTIONAL = più vicino ad 1

LESS\_FUNCTIONAL = più vicino a 0

ALTRI SCORE:

**integrated\_fitCons\_rankscore**: integrated fitCons scores were ranked among all integrated fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of integrated fitCons scores in dbNSFP.

**integrated\_confidence\_value**: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

**GM12878\_fitCons\_score**: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type GM12878.

**GM12878\_fitCons\_rankscore**: GM12878 fitCons scores were ranked among all GM12878 fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GM12878 fitCons scores in dbNSFP.

**GM12878\_confidence\_value**: 0 - highly significant scores (approx. p<.003); 1 - significant scores

(approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

**H1-hESC\_fitCons\_score**: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type H1-hESC.

**H1-hESC\_fitCons\_rankscore**: H1-hESC fitCons scores were ranked among all H1-hESC fitCons

scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of H1-hESC fitCons scores in dbNSFP.

**H1-hESC\_confidence\_value**: 0 - highly significant scores (approx. p<.003); 1 - significant scores

(approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

**HUVEC\_fitCons\_score**: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type HUVEC.

**HUVEC\_fitCons\_rankscore**: HUVEC fitCons scores were ranked among all HUVEC fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of HUVEC fitCons scores in dbNSFP.

**HUVEC\_confidence\_value**: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

**GERP++:**

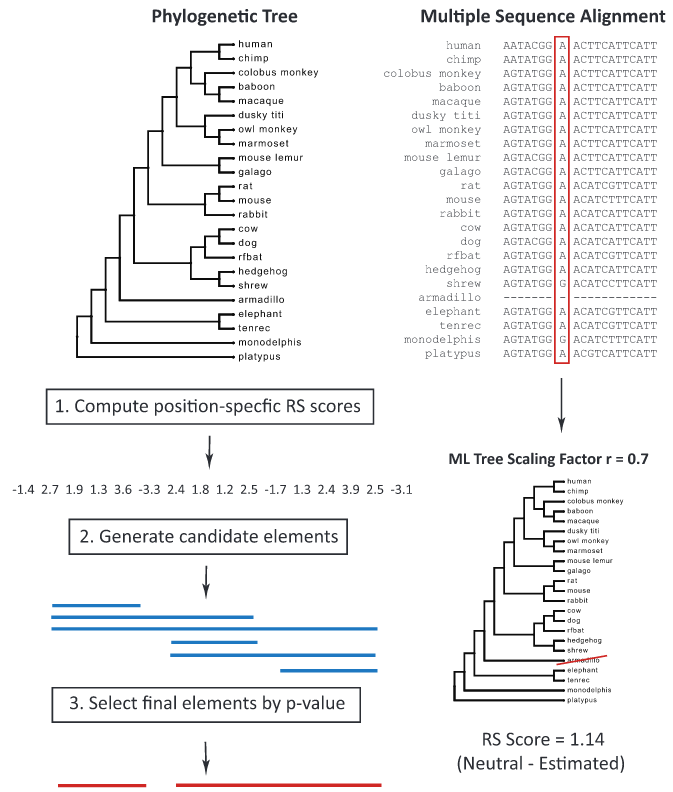
Il GERP++ è uno score di conservazione basato su singola base.

Stima il numero medio di sostituzioni in ogni genoma allineato per stimare una frequenza di evoluzione neutrale. Lo score originale è “Rejected Substitutions”: il numero di sostituzioni atteso sotto la condizioni di neutralità meno il numero di sostituzioni osservate ad ogni posizione allineata. Nuovi score sono basati su un modello Machine Learning della frequenza di sostituzione per la base.

Gli step sono i seguenti (rappresentati poi in immagine):

1. Per ogni posizione degli allineamenti multipli su 33 specie mammifere si calcola lo score di conservazione nelle sostituzioni rigettate sottraendo la stima della frequenza evolutiva meno la frequenza neutrale. L afreq. Neutrale viene calcolata rimuovendo le specie che hanno gap in quella posizione dall’albero filogenetico e sommando la lunghezza del braccio del’albero risultante; la frequenza evolutiva viene stimata calcolando la massima likelihood ridimensionata per l’albero progettato.
2. Dati degli score di conservazione posizione-specifica, si genera un set di elementi candidati.
3. Per ogni elemento candidato si calcola un p-value per rappresentare la likelihood di osservazione di segmenti di uguale dimensione e score maggiori o uguali sotto la condizione di modello nullo. Infine si seleziona un set di elementi non-overlappanti per aumentare il p-value.

L’immagine degli step viene di seguito riportata:



PREDIZIONE:

Lo score (GERP++\_RS) varia tra -12.3 e 6.17. Più lo score è alto, più la base è conservata.

ALTRI SCORE:

**GERP++\_NR**: GERP++ neutral rate

**GERP++\_RS\_rankscore**: GERP++ RS scores were ranked among all GERP++ RS scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GERP++ RS scores in dbNSFP.

**PHYLOP:**

Anche questo è uno score di conservazione basato su singola base.

Stima il numero medio di sostituzioni in ogni genoma allineato per stimare la frequenza evolutiva neutrale ricavata da dati non codificanti. Confronta la probabilità delle sostituzioni osservate sotto l’ipotesi di frequenza evolutiva neutrale.

Lo score si ottiene allineando un diverso numero di genomi. Ad esempio, PhyloP100way\_vertebrate è lo score ottenuto allineando i genomi di 100 vertebrati mentre il PhyloP20way è lo score ottenuto dall’allineamento di 20 vertebrati. Lo score viene ottenuto calcolando un p-value. Altri score poi sono disponibili per soli mammiferi, ecc...

PREDIZIONE:

Anche in questo caso, maggiore è lo score, più la sequenza è conservata e viceversa. Lo score varia in dbNSFP tra -20.0 e 10.003

ALTRI SCORE:

**phyloP100way\_vertebrate**: phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 100 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from -20.0 to 10.003 in dbNSFP.

**phyloP100way\_vertebrate\_rankscore**: phyloP100way\_vertebrate scores were ranked among all phyloP100way\_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP100way\_vertebrate scores in dbNSFP.

**phyloP20way\_mammalian**: phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from -13.282 to 1.199 in dbNSFP.

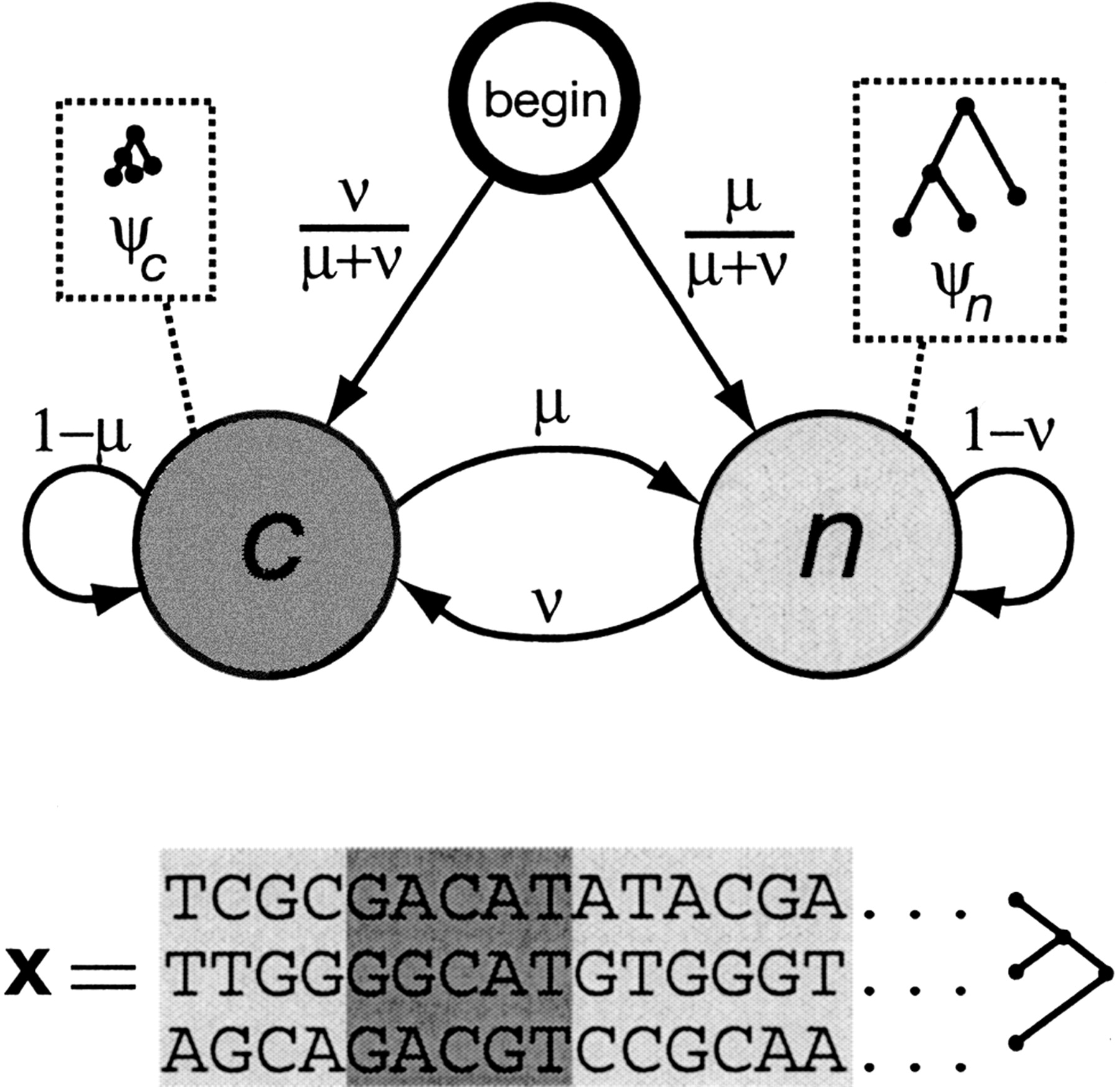
**phyloP20way\_mammalian\_rankscore**: phyloP20way\_mammalian scores were ranked among all phyloP20way\_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP20way\_mammalian scores in dbNSFP.

**PHASTCONS:**

PhastCons è uno score ottenuto dal software PhastCons che offre uno score di conservazione su piccole regioni in un multiplo allineamento, dato un albero filogenetico. È basato un modello phylo-HMM.

PhastCons fitta un Hidden Markov model (HMM) con gli stati “conservato” e “non conservato”. Le frequenze di sostituzione stimate sono ottenute allo stesso modo dello score PhyloP (vedi sopra). Il parametro u (mi greco) rappresenta l’inverso della lunghezza attesa delle regioni conservate.

Il parametro v (ni greco) setta la proporzione di regioni conservate:



PREDIZIONE:

Anche in questo, più lo score è alto, più è conservato. Lo score varia tra 0 e 1

E anche qui, si hanno diversi score, su allineamenti derivati da 100 vertebrati, 20 mammiferi, ecc...

ALTRI SCORE:

**phastCons100way\_vertebrate**: phastCons conservation score based on the multiple alignments of 100 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.

**phastCons100way\_vertebrate\_rankscore**: phastCons100way\_vertebrate scores were ranked among all phastCons100way\_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons100way\_vertebrate scores in dbNSFP.

**phastCons20way\_mammalian**: phastCons conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.

**phastCons20way\_mammalian\_rankscore**: phastCons20way\_mammalian scores were ranked among all phastCons20way\_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons20way\_mammalian scores in dbNSFP.

**SIPHY:**

SiPhy è uno score che guarda piccole regioni restituendo un modello di pattern di sostituzioni, piuttosto che solo la frequenza come fanno la maggior parte degli altri scores.

I modelli di sola frequenza hanno infatti due svantaggiprincipali:

* Non catturano tutti gli aspetti evolutivi di sequenze funzionali. I vincoli funzionali non devono semplicemente basarsi sulla diminuzione della frequenza di mutazione ma anche agire come un pattern di sostituzione affetto da bias. Ad esempio, si consideri una ridondante sostituzione aminoacidica. Molte basi possono liberamente cambiare fino a che non compromettono la sequenza proteica. Ad esempio, la Lysina è codificata da due codoni, AAA e AAG; il terzo nucleotide è libero di mutare tra A e G. Quindi, possono accadere svariate mutazioni tra A e G nel corso evolutivo. I metodi su frequenza la classificherebbero come non conservativa, mentre comunque il tratto può essere conservato dato che la sostituzione in questo caso porta alla stessa proteina.
* I metodi su frequenza riportano soltanto se una sequenza è conservata o meno, ma non forniscono informazioni in merito allo specifico constraint su ogni base singolarmente. Nell’esmpio della lysina, è più informativo dire che non solo la terza posizione è constrained, ma anche che che la base deve essere solo una purina.

Questo score identifica prima i siti sotto constraint come quelli con un pattern di sostituzione nucleotidica che varia significativamente dal pattern neutrale (si fa comunque un allineamento multiplo di sequenze). Nel caso del codone descritto prima, il pattern di sostituzione corrispondente alla terza posizione del codone sarà altamente pesata attraverso le sostituzioni A←→ G dal momento che tutte le altre sostituzioni cambiano l’aminoacido codificato.

PREDIZIONE:

Più lo score è grande, più sarà conservato. Lo score varia da 0 a 37.97

ALTRI SCORE:

**SiPhy\_29way\_pi**: The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes.

**SiPhy\_29way\_logOdds**: SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site. Scores range from 0 to 37.9718 in dbNSFP.

**SiPhy\_29way\_logOdds\_rankscore**: SiPhy\_29way\_logOdds scores were ranked among all SiPhy\_29way\_logOdds scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of SiPhy\_29way\_logOdds scores in dbNSFP.

**CLINVAR:**

ClinVar è un db in cui sono contenute informazioni in merito alle relazioni tra varianti e fenotipi con a supporto evidenze cliniche.

PREDIZIONE (non è una predizione ovviamente):

**clinvar\_clnsig:**

2 – Benign

3 - Likely benign

4 - Likely pathogenic

5 – Pathogenic

6 - drug response related

7 – histocompatibility related.

255 – other (risk factor, association, protective, etc...)

Uno score negativo indica che lo score è riferito all’allele di reference.

**Mendelian Clinically Applicable Pathogenicity (M-CAP) Score:**

M-CAP is the first pathogenicity classifier for rare missense variants in the human genome that is tuned to the high sensitivity required in the clinic (see Table). By combining previous pathogenicity scores (including SIFT, Polyphen-2 and CADD) with novel features and a powerful model, we attain the best classifier at all thresholds, reducing a typical exome/genome rare (<1%) missense variant (VUS) list from 300 to 120, while never mistaking 95% of known pathogenic variants as benign.

**MaxEntScan**

(MaxENT score) to a 9-mer (5 splice site) or a 23-mer (3 splice site) sequence. The higher the score, the higher the probability that the sequence is a true splice site. Also, it can be argued that given two sequences of differing scores, the higher scoring sequence has a higher likelihood of being used.

PREDIZIONE:

Maggiore è lo score, maggiore è la probabilità che la variante sia splice

**GLOSSARIO**

**CONSTRAINT:**

Constraint. Il termine ‘constraint’ (vincolo), nel significato piú accettato, può essere identificato in ‘ciò che dirige altri tipi di cambiamenti o che impedisce quei cambiamenti che sarebbero operati dalla selezione’; in senso lato, il ‘constraint’ è ‘un fattore che costringe o canalizza i cambiamenti fenotipici in una direzione stabilita dalla storia passata o dalla struttura formale, anziché dal corrente adattamento’ (Gould, 1989).  
E’ possibile individuare le seguenti categorie di constraint:  
(a) strutturale;  
(b) di sviluppo;  
(c) filogenetica;  
(d) genetica;  
(e) epigenetica;  
(f) comportamentale;  
(g) ecologica o ambientale;  
(h) selettiva.  
Uno dei piú importanti aspetti potrebbe identificarsi con il meccanismo di ‘assimilazione genetica’ (27), processo fondamentale della biologia epigenetica dello sviluppo, che Waddington (1957, 1975) ha evidenziato in Drosophila. La produzione di un nuovo assetto genetico, dovuto all’ ‘assimilazione genetica’ del nuovo fenotipo, è favorita dalla cosiddetta ‘selezione naturale’ nel corso dell’evoluzione ma, come nel caso dell’affine effetto Baldwin[31] (1986), il fenomeno è ‘canalizzato’ cioè guidato da precisi stimoli ambientali (Sarà, 2002)

**CONSTRAINTED GENETIC ELEMENTS:**

Nello stretto significato genetico, regioni “vincolate” sono quelle sequenze/regioni nel genoma che possono essere trovate preservate in svariate generazioni di una specie. Sono fattori che rendono le popolazioni resistenti ai cambi evoluzionari. Una definizione accettata è la seguente: “La proprietà di un tratto che, sebbene possibilmente adattativo nell’ambiente nel quale originariamente si è evoluto, agisce per porre dei limiti sulla produzione di nuove varianti fenotipiche.”

**FINGERPRINT**:

In biologia molecolare, si definisce **impronta genetica** (o **fingerprint**, o **fingerprinting genetico**) il profilo derivato dalla applicazione di determinati marcatori molecolari ad un genoma, al fine di renderlo riconoscibile e rintracciabile. Un marcatore molecolare (genetico) può evidenziare le differenze tra due genomi, in termini di lunghezza di regioni dalle estremità conosciute, o di presenza/assenza di una sequenza nota. La ricorrenza di determinate caratteristiche in genomi diversi può essere usata per definire il grado di imparentamento tra gruppi di organismi.Il fingerprinting genetico è stato introdotto per la prima volta da Alec Jeffreys nel 1985.

In **EPIGENOMICA** per fingerprint si intendono quelle posizioni genomiche appartenenti a specifiche classi funzionali.