Glossario

ACCESSIBILITÀ AL SOLVENTE L'area di superficie (solitamente misurata in angstrom quadri) di una molecola biologica che è esposta al solvente quando la molecola è avvolta nella sua forma tridimensionale nativa.

ACCESSION NUMBER Codice di accesso. È un identificativo fornito dai gestori delle più importanti banche dati biologiche in seguito a sottomissione di una nuova entry. L'accession number identifica ciascuna entry in modo univoco. Esempio: l'accession number della proteina RAS umana nella banca dati SWISSPROT è P01112.

ALBERO FILOGENETICO Rappresentazione delle relazioni fra gruppi tassonomici di organismi. Può essere radicato, quando i rami originano da un progenitore comune a tutti i gruppi tassonomici, oppure non radicato. Vedere anche Unità tassonomiche operative (OTU).

ALFABETO L'insieme dei simboli presenti in una biosequenza: le 4 basi (A, C, G, T) per le sequenze di DNA o le 4 basi (A, C, G, U) per l'RNA, e i 20 aminoacidi per le sequenze proteiche.

ALFANUMERICO Termine generico che indica i simboli usati per rappresentare, organizzare o controllare le informazioni. I caratteri alfanumerici comprendono le lettere dell'alfabeto (A-Z), le cifre arabe (0-9), i segni di interpunzione e altri segni specifici come la rappresentazione del dollaro e dell'euro.

ALGORITMO È sinonimo del termine "algorismo", ormai in disuso, che deriva dal nome del famoso matematico arabo del IX secolo Al Khuwarizmi; nel corso del XVII secolo il termine è cambiato in "algoritmo", incorporando la parola greca *arithmos* (numero). Con questo termine si indica una procedura formalizzata che permette di risolvere un problema in modo efficace, mediante sequenze ordinate e finite di operazioni elementari. Può essere rappresentato schematicamente sotto forma di flowchart o diagramma di flusso. I linguaggi di programmazione sono strumenti per l'implementazione di algoritmi.

ALLELE Forma alternativa di uno specifico *locus* (un singolo nucleotide polimorfico, un gene ecc.) nel genoma.

ALLINEAMENTO È il prodotto della procedura di confronto tra due o più sequenze in base all'ordine dei caratteri nelle sequenze in esame. Un allineamento può essere locale o globale: generalmente l'allineamento locale è il più utile dal punto di vista del biologo (Capitoli 3 e 4). Il migliore allineamento fra due sequenze è quello cui viene associato il più alto punteggio di similarità, così come determinato in base ai parametri prescelti (matrici di sostituzione e penalità associate ai gap).

ANALOGIA Relazione fra entità biologiche (microscopiche come proteine o macroscopiche come organi e apparati) che in specie diverse ricoprono lo stesso ruolo funzionale, senza avere un'origine evolutiva comune, risultati da evoluzione convergente. È l'opposto di omologia.

ANNOTAZIONE Una combinazione di commenti, note, referenze e citazioni, che descrivono tutta l'informazione disponibile su un gene, una proteina o altri sistemi biologici. Le annotazioni possono essere in formato libero o estratte da un vocabolario controllato (un elenco di termini alternativi e prefissati). L'annotazione automatica di sequenze o strutture biologiche a funzione ignota è una delle funzioni principali della bioinformatica.

ANTIGENE Qualsiasi molecola estranea capace di stimolare una risposta immunitaria in un vertebrato. Molti antigeni sono proteine, come per esempio le proteine dei capsidi virali.

ANTISENSO DNA/RNA con sequenza complementare a una data sequenza di DNA/RNA (detta invece *senso*). Possono avere ruoli di regolazione, quando per esempio un RNA antisenso si lega a un RNA messaggero "senso" in base alla complementarità di sequenza, impedendone la traduzione. Se ne fa un generico uso terapeutico, in quanto l'antisenso per specifiche sequenze di DNA o

RNA implicate in malattie si lega fisicamente a tali sequenze e ne inibisce l'espressione.

APLOIDE Una cellula o un organismo contenente un solo set di cromosomi (vedere anche diploide).

APLOTIPO Un insieme di alleli in specifici *loci* genomici che sono spesso ereditati insieme, perché raramente separati da un evento di ricombinazione. Vedere anche allele.

APPRENDIMENTO AUTOMATICO Vedere machine learning

ARCO Un elemento di un grafo che connette due nodi. L'arco può avere un verso (arco orientato) oppure non averlo (arco non orientato).

ASCII (*American Standard Code for Information Interchange*) Codice standard americano a 7 bit per l'interscambio di informazioni fra sistemi di elaborazione e di comunicazione. È in grado di codificare 128 (2^7) simboli diversi: di questi 96 sono caratteri alfanumerici, di punteggiatura e grafici; i restanti 32 sono caratteri di controllo delle periferiche e degli apparecchi di comunicazione. Per esempio, il carattere «A» è rappresentato dal codice 65, il carattere «3» dal codice 51 e la virgola dal codice 105.

ASSEMBLAGGIO Processo di ricostruzione della sequenza completa di un acido nucleico a partire dalle sequenze di corti frammenti.

BAC (*Bacterial Artificial Chromosome*) Vettore di clonaggio che può incorporare lunghi frammenti di DNA (100-300 kbp). Vedere anche YAC.

BAM Versione binaria del formato di interscambio SAM (Sequence Alignment/Map) per riportare le coordinate degli allineamenti di sequenze nucleotidiche.

BARCODING Metodo per l'identificazione di specie biologiche attraverso l'analisi di sequenze di DNA come dei marcatori (da *barcode*, codice a barre). Il marcatore deve avere sequenze diverse in specie diverse. Tra gli animali la sequenza scelta è quella del gene COX1 (la subunità 1 della citocromo ossidasi). Vedere anche DNA metabarcoding.

BASE CALLING Identificazione della sequenza nucleotidica fatta da software associati a piattaforme di sequenziamento, che convertono i segnali chimici/fisici prodotti dalla piattaforma in una sequenza.

BATTERIOFAGO Abbreviato anche come "fago", è un virus che infetta i batteri. Il DNA fagico è anche utilizzato come vettore di clonaggio e per la generazione di librerie peptidiche.

BED (*Browser Extensible Data*) Formato di file di interscambio per annotazioni genomiche. Riporta, fra le altre cose, le coordinate di ogni annotazione sotto forma di cromosoma, strand, posizione di inizio e fine dell'annotazione (per es. un gene).

BIOINFORMATICA Scienza interdisciplinare che coinvolge la biologia, l'informatica, la matematica e la statistica per l'analisi di sequenze biologiche, genomi e per la predizione della funzione e della struttura di macromolecole.

BIOSEQUENZA Successione di nucleotidi nel DNA e nel RNA, o aminoacidi nelle proteine, tenuti insieme da legami chimici e ordinati secondo regole che determinano l'espletamento di funzioni biologiche.

BIT (*BInary digiT*) Cifra binaria. Unità elementare di informazione del sistema binario, capace di assumere solo 2 valori, [0] e [1], corrispondenti rispettivamente allo stato della corrente elettrica di spento [off] e di acceso [on]. Un insieme di 8 bit costituisce il Byte, utilizzato per rappresentare qualsiasi tipo di informazione come le lettere dell'alfabeto [A-Z] e le cifre da 0 a 9.

BLAST (*Basic Local Alignment Search Tool*) Algoritmo per ricerche rapide di similarità in banche dati di biosequenze, partendo da una sequenza query. Ne esistono varie versioni, specializzate per ricerca di similarità fra proteine, sequenze nucleotidiche, traduzioni di acidi nucleici, e così via.

BLOSUM Vedere matrici di sostituzione.

BP (*base pair*) Paia di basi. Una coppia di nucleotidi le cui basi azotate (una purina e una pirimidina) sono unite da legami idrogeno: A:T e G:C nel DNA; A:U e G:C nell'RNA.

BROWSER Un programma client (in alternativa a server) che consente di accedere a documenti via Internet. I documenti sono accessibili in lettura. L'uso dei browser di rete è equivalente a un ftp anonimo. I browser più noti e utilizzati sono: Chrome, Edge, Firefox.

BYTE Unità di informazione, costituita da una successione di 8 bit, utilizzata per rappresentare nella CPU una entità logica (un carattere alfanumerico). Può assumere 256 (2^8) valori o combinazioni diverse. Il termine fu coniato nel 1964 dalla IBM.

C, LINGUAGGIO (*C language*) Linguaggio di programmazione di utilizzo relativamente difficile, ma che

consente di risolvere in modo efficace problemi molto complessi. È uno dei linguaggi di programmazione più diffusi.

C++, LINGUAGGIO (*C++ language*) Linguaggio di programmazione derivato da C, particolarmente indicato per la programmazione a oggetti (*Object oriented programming*).

cDNA (*complementary DNA*) Un filamento di DNA copiato sullo stampo dell'RNA messaggero maturo e quindi privo di introni. Vedere anche libreria di cDNA.

CDS (*coding regions*) Codice utilizzato nelle banche dati biologiche per descrivere la porzione di una sequenza genica che codifica una proteina (vedere anche Box 2.1).

CHIP (scheggia, frammento) Piastrina, normalmente di silicio, contenente un microcircuito elettronico integrato completo, capace di svolgere funzioni di memoria e logiche. L'attività del chip è regolata da un clock, con frequenza generalmente espressa in megahertz o gigahertz, che ha lo scopo di sincronizzare tutti i processi.

CHIP-SEQ Tecnologia basata su sequenziamento di frammenti nucleotidici immunoprecipitati usando anticorpi specifici per una data proteina interagente con il DNA genomico.

CIGAR Formato compatto per descrivere l'allineamento di una sequenza nucleotidica corta con una sequenza di riferimento. Le stringhe CIGAR sono incluse nel formato di interscambio SAM/BAM.

CLIENT Un computer (o un software) remoto per connettersi e ottenere dati da un computer (o da un software) server. Un browser web è un software client in grado di comunicare con un computer remoto sul quale funzioni l'appropriato web server.

CLOCK (orologio) Dispositivo elettronico, basato su un cristallo di quarzo, che genera impulsi a intervalli regolari allo scopo di sincronizzare le operazioni del processore di un computer. La velocità del clock, espressa in megahertz o gigahertz, è un indicatore della potenza di un personal computer.

CLONAGGIO Termine che indica sia la generazione di cloni o di repliche genetiche esatte, sia l'inserzione di un DNA esogeno in un vettore.

CLONE Una popolazione di molecole di DNA o di cellule geneticamente identiche.

CLUSTERING (raggruppamento) Procedura algoritmica per dividere un insieme di dati in gruppi (cluster) tali che la distanza intra-cluster (cioè fra membri dello stesso gruppo) sia minimizzata, e che la distanza inter-cluster (cioè fra membri di gruppi diversi) sia massimizzata. È un esempio di machine learning non supervisionato.

CNV (*Copy Number Variation*) Variazione del numero di copie di una porzione del genoma rispetto a un riferimento. Può essere causata da una delezione o da un'amplificazione.

CODICE GENETICO La tabella di corrispondenza tra triplette di nucleotidi e aminoacidi: le 64 possibili triplette codificano i 20 aminoacidi che si trovano nelle proteine biologiche, oltre ai codoni di terminazione (che segnalano all'apparato di traduzione la fine della catena proteica). Il codice genetico è ridondante e degenerato.

COEFFICIENTE DI CORRELAZIONE Una misura del grado di relazione lineare tra due variabili, il cui valore può variare tra –1 e +1. Un valore positivo indica una proporzionalità diretta, un valore negativo indica una proporzionalità inversa tra le due variabili. La distanza del valore dallo zero indica la forza della relazione.

COG (*Cluster of Orthologous Groups*) Gruppi di sequenze evolutivamente correlate, ma appartenenti a genomi diversi. Vedere anche ortologia.

COMANDO Una o più parole specifiche, definite nei sistemi operativi o nei menu di gestione dei programmi, che, se digitate tramite la tastiera o attivate mediante il mouse, eseguono un'operazione finalizzata a ottenere un risultato. Per esempio, il comando "ls" nel sistema operativo Unix consente di ottenere la lista dei file in una directory.

COMPILATORE Programma la cui funzione è convertire un programma scritto con un linguaggio di alto livello in un oggetto codificato nel linguaggio macchina e quindi interpretabile da parte del computer.

COMPLESSITÀ (di un algoritmo) Descrive il numero di passi impiegati in un algoritmo per risolvere un problema come funzione della quantità di dati, per esempio la lunghezza delle sequenze da allineare.

COMPLESSITÀ (di una biosequenza) Misura del contenuto informazionale di una biosequenza. Per esempio, biosequenze a bassa complessità mostrano una composizione dei nucleotidi o degli aminoacidi fortemente sbilanciata o ripetitiva.

CONFORMAZIONE La precisa posizione tridimensionale di atomi e legami in una molecola.

CONSENSUS Sequenza che rappresenta un allineamento multiplo. La sequenza *consensus* riporta nelle posizioni variabili i residui possibili alternativi secondo un simbolismo convenzionale.

CONTIG Un insieme di sequenze contigue che possono essere assemblate in un ordine lineare.

COOKIE Piccola quantità di informazione scambiata tra un server e un browser web e che viene conservata, per una quantità di tempo definita, sul computer client. Il browser può accettare o non accettare il cookie, a seconda delle preferenze dell'utente. In generale, i cookies contengono informazioni come: preferenze dell'utente, dati di registrazione e così via.

COVERAGE Parametro che riporta la "copertura" media di una sequenza di acido nucleico ricostruita da una procedura di assemblaggio, in base a quante sequenze (read) sono state utilizzate per ricostruirlo. Un coverage 10× indica che in media ogni nucleotide della sequenza ricostruita è supportato da 10 diverse read.

CPU (*Central Processing Unit*) Unità di memoria centrale del calcolatore dove ha sede l'elaborazione dei processi.

CROMATINA Il complesso costituito dal DNA e dalle proteine a esso associate (prevalentemente istoni) durante l'interfase.

CROMOSOMA Unità strutturale del materiale genetico eucariotico costituito da una molecola di DNA e dalle proteine a esso associate.

CROSSING-OVER Ricombinazione di cromosomi in fase replicativa durante la quale avviene lo scambio di materiale genetico fra i cromatidi materni e quelli paterni.

CURVA ROC (*Receiver Operating Characteristic*) Metodo per la valutazione della performance di un algoritmo predittivo, ottenuta riportando in un grafico la variazione del tasso di veri positivi identificati rispetto al tasso dei falsi positivi. L'area sottesa dalla curva (AUC) è un valore rappresentativo, che va da 1 (predittore perfetto) a 0,5 (predittore randomico).

DATA BASE (banca dati) Anche *data bank* (vedere Capitolo 2). Vasto insieme di informazioni relative a un settore di conoscenza o a un'organizzazione, strutturate in modo tale che siano possibili facilmente le operazioni di consultazione, ordinamento, aggiornamento e stampa di report personalizzati. I principali modelli di organizzazione di un database sono di tipo gerarchico, reticolare e relazionale.

DATA MINING Processo di estrazione di informazioni rilevanti dall'osservazione di quantità massive di dati. Il data mining può avvenire mediante l'applicazione di algoritmi computazionali e tecniche statistiche a testi e banche dati.

DEBUGGING Processo di ricerca ed eliminazione degli errori in un programma informatico.

DEFAULT (difetto) Valore di un parametro assunto in modo automatico dal computer, corrispondente alle situazioni più frequenti e utili, in mancanza (in difetto) di indicazioni specifiche da parte dell'operatore.

DELEZIONE Alterazione genetica in cui una porzione di DNA (da una singola base a una porzione di cromosoma) viene perduta.

DIPLOIDE Una cellula o un organismo contenente due set di cromosomi omologhi.

DIRECTORY Area di un disco in cui possono essere memorizzati i file. Il processo di formattazione di un disco genera automaticamente la directory principale (root directory). La directory può essere strutturata ad albero e contenere altre directory (subdirectory). L'insieme delle directory e delle subdirectory, presenti sul disco, è chiamata struttura della directory. Nei sistemi operativi user-friendly dei personal computer, le directory appaiono come cartelle o folder.

DNA METABARCODING Metodo per determinare la composizione in specie di una comunità (per es. un microbioma) in esperimenti di metagenomica. Invece di sequenziare tutto il DNA estratto dal campione, si amplifica per PCR e poi si sequenzia uno specifico marcatore (il *barcode*, codice a barre), scelto in modo tale che abbia una sequenza diversa in specie diverse. Esempi di barcode sono l'rRNA 16S o alcuni geni di mitocondri o cloroplasti. Vedere anche barcoding.

DNS (*Domain Name System*) È un insieme di dati, software e protocolli che consentono di tradurre il nome di un dominio come «www.ncbi.nlm.nih.gov» in un indirizzo IP come 130.14.29.110. La richiesta di una tale traduzione si definisce DNS query ed è effettuata dal sistema ogni qualvolta si voglia accedere a un dominio utilizzando il nome piuttosto che l'indirizzo IP.

DOCKING Procedura per determinare o verificare l'interazione fra due molecole, per esempio fra due proteine, fra una proteina e un acido nucleico, fra una proteina e una piccola molecola.

DOMINANTE Ogni carattere che viene espresso fenotipicamente anche in presenza di alleli che, se presenti su

entrambi i cromosomi, darebbero un fenotipo diverso. Vedere anche recessivo.

DOMINIO Indirizzo logico o nome che identifica un computer della rete ed è generalmente composto di almeno due parti, separate da un punto (per esempio cbm.bio.uniroma2.it). Il nome a sinistra è il più specifico. La parte rimanente del nome è comune a tutti i computer in una LAN.

ELETTROFORESI Uso di un campo elettrico per separare frammenti di biomolecole (acidi nucleici o proteine) sulla base della loro carica e della loro conformazione in gel di agarosio o di acrilammide.

ENTRY Entità della banca dati che racchiude il set di informazioni associate a ciascuna unità elementare della banca dati (per es. l'entry RASH_HUMAN della banca dati SWISSPROT contiene informazioni di tipo funzionale e bibliografico sulla proteina umana RAS e la sequenza stessa).

ENZIMI DI RESTRIZIONE Un tipo di enzima che riconosce corte sequenze di DNA (solitamente palindromiche) e le taglia su entrambe le eliche. Sono le forbici molecolari che consentono di operare la tecnologia del DNA ricombinante.

ENZIMI Una classe di proteine in grado di catalizzare reazioni chimiche (per es. la formazione o la rottura di un legame chimico) che sarebbero altrimenti irrealizzabili in condizioni fisiologiche.

ESOMA Parte del DNA genomico corrispondente ai soli esoni dei geni. Vedere anche WES.

ESONE Regione genica che contiene parte dell'informazione codificante per una proteina o per un RNA non codificante. Nell'RNA messaggero sono anche presenti esoni o parti di esoni non codificanti corrispondenti alle regioni non tradotte dell'mRNA (5'UTR e 3'UTR). Tipicamente una proteina eucariotica è composta da diversi domini codificati da diversi esoni. Vedere anche introni.

ESPRESSIONE (genica) Misura della presenza di uno o più prodotti genici in un particolare tipo cellulare o in un tessuto. Studi di espressione si effettuano a livello dell'RNA o delle proteine allo scopo di determinare il numero, il tipo e il livello dei prodotti genici che possono essere regolati durante il ciclo cellulare, in risposta a stimoli esterni o in dipendenza da malattie. La trascrittomica e la proteomica ora consentono lo studio dei profili di espressione (*expression profile*) di insiemi di geni o di interi genomi.

ESPRESSIONE REGOLARE (*regular expression*, regexp) Stringa di simboli che descrive un insieme di stringhe, riportandone la variabilità in forma compatta e secondo una serie di convenzioni.

EST (*Expressed Sequence Tags*) Una corta sequenza proveniente da un gene espresso. Tipicamente le sequenze EST vengono identificate attraverso la purificazione dell'RNA messaggero, la conversione in cDNA e il successivo sequenziamento degli inserti di cDNA presenti nei cloni corrispondenti a una o ad entrambe le estremità.

ETERODIMERO Proteina composta da due diverse catene o subunità.

ETEROPLASMIA Presenza di genomi differenti in diversi mitocondri di una cellula. Il contrario di omoplasmia.

ETHERNET Un mezzo di connessione per computer appartenenti a una LAN.

EUCARIOTI Una cellula o un organismo composto da cellule con un nucleo circondato da membrana (vedere procarioti).

EURISTICA (o algoritmo euristico) Tipo di algoritmo che non garantisce di trovare la soluzione ottimale di un problema, ma, in base ad assunzioni e approssimazioni, permette di trovare soluzioni possibilmente valide in tempi rapidi. Si usano quando algoritmi esatti sono troppo complessi o non sono disponibili per risolvere un dato problema (vedere NP-HARD).

FATTORI DI TRASCRIZIONE Proteine regolatrici che si legano alle sequenze del promotore e sono necessarie per la trascrizione di un gene.

FENOTIPO Ogni caratteristica manifestata da un determinato carattere genetico, qualitativa o quantitativa (vedere anche genotipo).

FILE (archivio) Insieme ordinato di informazioni che possono essere memorizzate o elaborate in modo unitario. Ogni file è identificato da un nome, le cui regole per l'assegnazione dipendono dai sistemi operativi.

FORMATO FASTA Una sequenza in formato FASTA comincia con una singola riga di descrizione in cui il primo carattere è «>», seguita dalle righe contenenti la sequenza vera e propria. In generale, le righe devono essere di lunghezza minore di 80 caratteri. I nucleotidi sono identificati coi caratteri G, A, T o U, C. Gli aminoacidi sono descritti con il codice a una lettera (A: Alanina e così via); i caratteri non identificati vengono descritti con una X (Figura 7.8).

© 978-88-08-62112-2

FORMATO FASTQ Formato di file utilizzato per riportare le sequenze lette da un sequenziatore. Riporta, oltre alla sequenza nucleotidica, anche dei punteggi di qualità (PHRED) associati a ogni nucleotide letto e calcolati in base a una stima della probabilità che la lettura sia errata.

FORMATO Molti programmi richiedono che i dati vengano specificati in modo formalmente prefissato, usando particolari termini e un determinato ordine dei dati. Queste particolari specifiche si identificano col formato dei file.

FRAMESHIFT Una delezione, sostituzione o duplicazione di una o più basi nel DNA che provoca uno spostamento della corretta fase di lettura a triplette.

FTP (*File Transfer Protocol*) Protocollo per il trasferimento di dati tra computer remoti, che può avvenire in due modalità: con username e password (quando si possiedono privilegi di lettura e scrittura nel computer remoto) oppure anonima (in sola lettura, per esempio per scaricare dati). Nel caso di ftp anonimo, vengono comunque richiesti uno username (bisogna utilizzare lo username anonymous) e una password (in generale il proprio indirizzo di posta elettronica).

GAP Disallineamento tra due biosequenze causato da un'inserzione o una delezione in una delle due sequenze. Un gap può essere lungo da 1 a *n* nucleotidi o aminoacidi. Nel calcolo del punteggio da associare a un allineamento tra sequenze, i gap causano un decremento del punteggio (gap penalty) che può venire misurato in diversi modi.

GAP PENALTY La penalizzazione apportata al punteggio di similarità e dovuta a un'inserzione o delezione (gap), alla lunghezza dell'inserzione/delezione, o a entrambi i fattori.

GENBANK Banca dati di sequenze genetiche organizzata e mantenuta presso l'NIH (USA).

GENOMA Il materiale genetico di un organismo.

GENOME BROWSER Interfaccia, accessibile tramite un internet browser, di alcune banche dati genomiche, che permette la visualizzazione interattiva delle sequenze genomiche e delle loro annotazioni.

GENOMICA Analisi dell'intero genoma di un organismo.

GENOMICA FUNZIONALE Assegnazione della funzione genica attraverso il confronto tra genomi o specifiche indagini sperimentali (per es. lo studio del fenotipo mutante).

GENOTIPO Il corredo genetico di un individuo. Il termine può essere riferito al genoma nel suo insieme oppure a un *locus* specifico.

GFF (*General Feature Format*) Formato di file di interscambio di annotazioni genomiche, più esteso del formato BED. La variante del formato più utilizzata è la versione 2, detta anche GTF (*Gene Transfer Format*).

GIGABYTE Unità di misura della memoria di un computer. Equivale a 1 073 741 824 (230) byte e a 1024 megabyte.

GRAFO Struttura matematica formata da un insieme di nodi (o vertici) che possono essere collegati fra loro da archi (o lati). Gli archi possono essere orientati, cioè avere una direzione. I grafi sono molto usati in molte applicazioni bioinformatiche, per esempio per descrivere le reti di interazioni fra molecole.

GTF (*Gene Transfer Format*) Formato di file di interscambio per annotazioni genomiche. È una versione specifica del formato GFF.

GWAS (*Genome Wide Association Studies*). Metodi per determinare la significativa associazione fra varianti genetiche e tratti fenotipici, per esempio patologici.

HARDWARE Insieme dei dispositivi meccanici, elettrici, magnetici, elettronici (chip, schede, cavi, memorie, periferiche ecc.) costituenti un sistema per l'elaborazione elettronica dei dati. Si riferisce all'aspetto fisico del computer in contrapposizione al termine software.

HDF5 (*Hierarchical Data Format* 5) Formato di file prodotto in output da sequenziatori Pacific Biosystems e Oxford Nanopore, caratterizzato da una struttura gerarchica e ricca di informazioni sulle sequenze lette.

HIGH-THROUGHPUT Metodologia attraverso la quale un grandissimo numero di composti (geni, o proteine) viene analizzato, in genere con l'aiuto di metodi automatici o robotizzati.

HMM (*Hidden Markov Model*) Modello statistico usato per descrivere dipendenze fra entità, in cui il sistema è modellato come un processo markoviano. È composto da una serie di nodi (stati), connessi fra loro da archi, con associate delle probabilità, dette di transizione, di passare da uno stato all'altro. Ad ogni stato possono poi essere associate delle probabilità dette di emissione, che descrivono le probabilità di eventi osservabili. Data una sequenza di osservazioni, si cerca di determinare il più

probabile percorso fra gli stati del modello che può averne dato luogo.

IBRIDIZZAZIONE L'interazione tra filamenti complementari di acidi nucleici, che può essere messa in evidenza attraverso l'applicazione di diversi approcci sperimentali.

IDENTITÀ Misura che si può associare a un allineamento tra sequenze nucleotidiche o proteiche e che si ottiene contando le coppie di residui identici. In genere a un allineamento si associa la percentuale di identità tra le due sequenze (vedere anche similarità).

IN SILICO (anche usato *in silicio*) Letteralmente: all'interno di un computer. L'uso di metodi computazionali per simulare processi o analizzare un esperimento biologico. Si usa in alternativa a *in vivo* e *in vitro*.

INDEL Inserzione o delezione in un allineamento di sequenze.

INDIRIZZO IP Vedere IP.

INGEGNERIA GENETICA (o tecnologia del DNA ricombinante) È la procedura per isolare e manipolare il DNA al di fuori della cellula. Consente a un frammento di DNA di essere introdotto in una cellula od organismo diverso da quello d'origine, di replicarsi ed esprimersi.

INTERATTOMA L'insieme delle interazioni fra molecole in un organismo.

INTERNET La rete mondiale di computer collegati attraverso il protocollo TCP/IP.

INTRONI Sequenze geniche di solito trovate negli eucarioti che separano le regioni dei geni (esoni) che faranno parte di un trascritto maturo. Le sequenze degli introni vengono eliminate dai trascritti di RNA attraverso il processo dello splicing.

IP (*Internet Protocol*) È il protocollo che utilizza gli indirizzi IP (IP address). Un indirizzo IP è formato da 4 numeri compresi da 0 a 255, separati da punti (esempio 130.14.29.110). Vedere anche DNS.

KILOBYTE Anche Kb o K. Multiplo dell'unità di misura della memoria del computer, equivale a 1024 byte (2^{10}).

LAN (*Local Area Network*) Rete locale di comunicazione tra computer che condividono una serie di risorse (hard disk di grande capacità, programmi, banche dati, stampanti laser) all'interno di un unico edificio o complesso di edifici limitrofi.

LEGAME FOSFODIESTERICO Legame covalente che si forma fra il carbonio in posizione 3' dello zucchero di un nucleotide e il carbonio in posizione 5' di un altro nucleotide, liberando un pirofosfato.

LEGAME IDROGENO Un legame chimico debole tra un atomo elettronegativo (per es. l'ossigeno) e un atomo di idrogeno covalentemente attaccato a un altro atomo elettronegativo. I legami idrogeno sono alla base delle interazioni tra le due eliche del DNA e tra i residui aminoacidici nelle strutture secondarie delle proteine.

LEGAME PEPTIDICO Legame covalente formato tra due aminoacidi quando il gruppo amminico dell'uno si unisce al gruppo carbossilico dell'altro con l'eliminazione di una molecola d'acqua. Il legame peptidico è un legame forte, con carattere di legame parzialmente doppio.

LIBRERIA DI cDNA L'insieme dei frammenti di DNA ottenuti da una preparazione di RNA messaggero in un determinato tipo cellulare, tessuto oppure organismo.

LIGANDO Ogni molecola che si lega a una proteina oppure a un recettore; il partner di interazione di proteine, enzimi o recettori.

LINGUAGGIO DI ALTO LIVELLO Linguaggio di programmazione evoluto (Basic, C, FORTRAN…) orientato al problema, vicino alla logica e al linguaggio umano, che utilizza comandi e simboli generalmente non dipendenti dal sistema operativo e dalle caratteristiche hardware del computer. I programmi, scritti con i linguaggi di alto livello, per essere eseguiti dal computer devono essere tradotti e resi eseguibili mediante i compilatori.

LINGUAGGIO DI BASSO LIVELLO Linguaggio di programmazione (Linguaggio macchina, Assembly language) orientato alla macchina che consente di scrivere programmi efficienti e compatti. Lontano dalla logica e dal linguaggio umano, è destinato agli specialisti.

LINGUAGGIO DI PROGRAMMAZIONE Linguaggio artificiale che consente di codificare i passi che il computer deve eseguire per svolgere un processo di gestione del computer stesso o un processo di elaborazione dei dati. Il linguaggio di programmazione, mediante precise regole e una speciale simbologia, è in grado di descrivere gli algoritmi, le azioni e gli oggetti. È caratterizzato da un alfabeto (segni grafici riconosciuti), da una grammatica (regole per il riconoscimento delle parole e per costruire frasi corrette) e da una semantica (assegnazione di significati alle stringhe). Esistono due categorie di linguaggi: i linguaggi di basso e di alto livello. I programmi, scritti con qualsiasi linguaggio, per funzionare devono comunque essere tradotti da un compilatore in codice binario, l'unico comprensibile alla macchina.

LINGUAGGIO DI SCRIPTING Linguaggio di alto livello che può essere usato per generare procedure dette script. Gli script non devono essere compilati (tradotti in codice binario), ma vengono eseguiti da un particolare programma detto interprete. L'esecuzione di uno script è generalmente più lenta dell'esecuzione di un programma compilato di pari complessità.

LINGUAGGIO MACCHINA Linguaggio di programmazione di basso livello, orientato alla macchina, che utilizza istruzioni binarie [0] e [1]. È l'unico linguaggio che l'unità centrale del computer (CPU) è in grado di interpretare, perciò i programmi scritti in linguaggio di alto livello, per essere compresi, devono essere tradotti da un compilatore in linguaggio macchina.

LINKAGE DISEQUILIBRIUM Associazione non casuale fra alleli in specifici *loci*, dovuta a bassa frequenza di ricombinazione, generalmente causata da vicinanza fra i *loci* nel genoma. Vedere anche aplotipo.

LINKAGE Misura della distanza di geni (o di *loci* genici) sullo stesso cromosoma basata sulla frequenza di ricombinazione osservata durante la meiosi.

LOCUS La specifica posizione occupata da un gene in un cromosoma. Diverse varianti di uno stesso gene (alleli) possono essere presenti in alternativa nello stesso *locus*.

MACHINE LEARNING (apprendimento automatico) Classe di algoritmi in grado di "apprendere" da un insieme di dati. Può essere supervisionato o non supervisionato (vedere Box 14.1).

MATRICE DI CONFUSIONE Matrice 2×2 che riporta il totale di veri positivi (VP), veri negativi (VN), falsi positivi (FP) e falsi negativi (FN) prodotti da un algoritmo predittivo su un dataset. Da questa matrice si possono calcolare varie metriche che misurano la performance predittiva dell'algoritmo.

MATRICE DI SOSTITUZIONE Matrice 20×20 contenente i coefficienti di similarità fra coppie di aminoacidi. Esistono differenti matrici di sostituzione (PAM, BLOSUM, Dayhoff ecc.) generate con algoritmi che stimano la probabilità di conversione di un aminoacido in un altro o la probabilità di conservazione di un aminoacido a partire da allineamenti multipli di sequenze proteiche evolutivamente correlate (Paragrafo 5.4). Tali matrici vengono utilizzate per associare dei punteggi agli allineamenti tra sequenze proteiche.

MATRICE POSIZIONALE DI PESO Vedere PWM.

MEGABYTE Anche M o Mb. Multiplo dell'unità di misura della memoria del computer, equivale a 1 048 576 byte (2^{20}).

MEMORIA CACHE (memoria nascosta) Area di memoria destinata a fare da ponte tra il processore e la memoria principale. L'uso della cache migliora di molto le prestazioni di un computer, ospitando dati di accesso frequente.

MEMORIA Componente fondamentale di un computer, indica qualsiasi dispositivo capace di conservare rappresentazioni codificate in forma binaria di programmi, dati ed elaborazioni. Esistono diversi tipi di memorie: RAM, ROM, memoria di massa. La capacità della memoria si misura in byte, kilobyte, megabyte, gigabyte e terabyte.

MEMORIA DI MASSA Anche memoria esterna. Qualsiasi dispositivo (hard disk, CD, DVD...) capace di memorizzare in modo permanente programmi, dati ed elaborazioni, successivamente richiamabili. La capacità della memoria di massa si misura in byte (gigabyte, terabyte). Quando il computer si spegne, la memoria di massa conserva le informazioni (diversamente da quanto succede per la RAM).

MEMORIA PRINCIPALE Anche RAM (*Random Access Memory*). Memoria primaria interna del computer, costituita da numerosi circuiti elettronici ad accesso casuale (cioè accessibile in qualsiasi punto della memoria), che contiene i programmi e i dati da elaborare. Tali programmi e dati possono essere immessi da un operatore, mediante digitazione dalla tastiera o per trasferimento dalla memoria di massa. La capacità della memoria centrale si misura in byte (megabyte, gigabyte).

MEMORIA VIRTUALE Espansione della memoria RAM ottenuta mediante una gestione simbiotica con la memoria di massa (hard disk). Consente ai programmi applicativi di funzionare come se il sistema disponesse di una quantità di memoria superiore a quella realmente installata. Quando si usa la memoria virtuale, i tempi di calcolo si allungano, poiché la memoria di massa ha tempi di risposta intrinsecamente più lenti di quelli della RAM.

METAGENOMICA Studio del genoma di una comunità, generalmente microbica, che vive nello stesso luogo (per es. un distretto corporeo, un ambiente ecologico). Vedere anche microbioma.

METILAZIONE L'aggiunta di un gruppo -CH_3 (metile) a un substrato, tipicamente il nucleotide citosina nel DNA.

MICROARRAY Tecnologia che permette il rilevamento e la stima dell'abbondanza di molecole presenti in un campione, basata su un supporto solido su cui sono coniugate delle sonde in grado di interagire con ciascuna di queste molecole. Nei microarray usati per misure dell'espressione genica, sulla superficie dell'array sono depositati o sintetizzati DNA (oligonucleotidi o interi cDNA) complementari a trascritti noti per una data specie.

MICROBIOMA Insieme del patrimonio genetico di una comunità microbica.

MICROSATELLITI Tratto di DNA altamente ripetitivo formato da un'unità (di solito lunga da 1 a 6 nucleotidi) ripetuta in tandem alcune decine di volte, e che può trovarsi in migliaia di siti diversi in un genoma. Polimorfismi nel numero di ripetizioni dell'unità di base sono usati in genomica forense.

MINISATELLITI Tratto di DNA altamente ripetitivo formato da un'unità (di solito lunga fino a qualche decina di nucleotidi) ripetuta in tandem alcune decine di volte, e che può trovarsi in migliaia di siti diversi in un genoma.

MODELING In bioinformatica, ci si riferisce in genere al modeling come alla tecnica che consente di inferire la struttura di una proteina a partire da informazioni sulla sua sequenza.

MODIFICAZIONI POST-TRADUZIONALI Modificazioni apportate a una proteina dopo la sua sintesi. Esempi di modificazioni post-traduzionali sono l'aggiunta di gruppi fosfato, acetile, carboidrati, lipidi ecc. che possono essere critici per la funzione della proteina.

MODIFICAZIONI POST-TRASCRIZIONALI Modificazioni apportate al pre-mRNA prima che lasci il nucleo e diventi RNA messaggero maturo.

MONOMERO Singola unità di una molecola o macromolecola biologica. Esempi di monomeri sono: singoli aminoacidi, domini o proteine.

MOTIVO Elemento conservato in un allineamento di sequenze nucleotidiche o proteiche che di solito si associa a una determinata funzione (Capitoli 8 e 13). Detto anche pattern.

MUTAGENO Agente che causa un aumento della frequenza delle mutazioni di un organismo.

MUTAZIONE Alterazione genetica ereditabile (mutazione germinale) o acquisibile dall'individuo nell'arco della sua vita (mutazione somatica), che include: mutazioni puntiformi (inserzioni, delezioni o mutazioni di senso riguardanti un unico nucleotide) o alterazioni su più larga scala, tra cui riarrangiamenti cromosomici.

N50 Misura della qualità della ricostruzione di una sequenza genomica, che stima la dimensione dei frammenti assemblati necessaria per rappresentare almeno la metà del genoma. Vedere anche assemblaggio.

NEEDLEMAN E WUNSCH, ALGORITMO DI Algoritmo di programmazione dinamica per l'allineamento globale di due biosequenze.

NGS (*Next Generation Sequencing*) Tecnologie di sequenziamento parallelo degli acidi nucleici ad alta resa capaci di produrre un enorme numero di letture (READ) di frammenti genomici estratti dalle cellule di un campione.

NODO Termine che indica: la porta di accesso a una rete telematica; elemento di un grafo, connesso ad altri nodi da archi; elemento di una struttura contenente informazioni che sono in relazione ad altre simili.

NP-HARD (Non deterministico Polinomiale difficile) Un problema per cui non esiste una soluzione algoritmica con complessità polinomiale. Per approcciare questi problemi si deve quindi ricorrere a euristiche.

OLC (*Overlap-Layout-Consensus*) Classe di algoritmi di assemblaggio per la ricostruzione di una sequenza genomica, basata sul confronto fra tutte le read disponibili, la loro organizzazione in gruppi contigui (CONTIG) e la derivazione di una sequenza consenso rappresentativa di ogni gruppo.

OLIGONUCLEOTIDE Corta molecola costituita da un numero relativamente basso di nucleotidi (tipicamente tra 10 e 60) uniti da legami fosfodiesterici.

OMICA Scienza che studia un particolare aspetto dei sistemi biologici (genoma, trascrittoma, proteoma, interattoma ecc.) in maniera globale.

OMOLOGIA Due geni o proteine si definiscono omologhi se si sono evoluti da un progenitore comune (vedere anche ortologia e paralogia). Talvolta, e non correttamente, si confondono i concetti di similarità e di omologia (vedere anche Capitolo 4).

OMOPLASMIA Presenza di genomi identici in tutti i mitocondri di una cellula. Il contrario di eteroplasmia.

OPERATORE Corta sequenza regolatrice di DNA che interagisce con il prodotto di un gene regolatore e con-

© 978-88-08-62112-2

trolla la trascrizione di uno o più geni costituenti un operone.

OPERONE Unità di trascrizione contenente uno o più geni strutturali, un promotore e un operatore. I geni di un operone vengono espressi in modo coordinato.

ORF (*Open Reading Frame*) Ogni frammento di DNA contenente una fase di lettura aperta (non interrotta da codoni di terminazione).

OROLOGIO MOLECOLARE Metodo per lo studio dell'evoluzione molecolare, basato sull'assunzione che mutazioni neutrali si accumulino nella sequenza di un gene o proteina con frequenza costante.

ORTOLOGIA Una coppia di geni appartenenti a due specie diverse si dice ortologa quando si presume che i due geni abbiano cominciato a divergere in seguito al processo di speciazione delle specie considerate. Proteine ortologhe svolgono generalmente la stessa funzione nei due organismi.

PACCHETTO SOFTWARE Prodotto software, costituito da un insieme di programmi fra loro correlati e dai relativi manuali, destinato a risolvere problemi applicativi.

PALINDROME Regioni di DNA o RNA con sequenze di basi complementari e invertite. Sequenze palindromiche possono formare particolari strutture (dette *hairpin*) nelle molecole degli acidi nucleici, talvolta con significato funzionale. Sequenze palindromiche sono quelle riconosciute dagli enzimi di restrizione.

PAM (*Percent Accepted Mutation*) Vedere matrici di sostituzione.

PARALOGIA Due geni si dicono paraloghi se derivano da un evento di duplicazione genica. Mentre geni ortologhi hanno spesso la stessa funzione, geni paraloghi in genere svolgono funzioni diverse, anche se spesso correlate.

PATTERN, vedi motivo.

PCR (*Polymerase Chain Reaction*) Tecnica utilizzata per amplificare o generare grandi quantitativi di copie di un frammento di DNA di origine qualsiasi. Condizione necessaria per poter applicare tale tecnica è la conoscenza della sequenza delle estremità della molecola da amplificare (*primer*).

PEAK CALLER Metodo di analisi della distribuzione di sequenze NGS ottenute da un campione e mappate su una sequenza genomica di riferimento, allo scopo di identificare regioni discrete e valutarne la significatività. Spesso usati per analisi di esperimenti di ChIP-Seq o simili.

PENETRANZA Frequenza con cui un allele mostra il fenotipo a esso associato. Nel caso di fenotipi patologici, mostra la frequenza con cui individui che portano l'allele nel loro genotipo svilupperanno la patologia.

PERCORSO EULERIANO (o cammino Euleriano) Percorso in un grafo che passa per tutti gli archi del grafo esattamente una volta.

PERCORSO HAMILTONIANO (o cammino Hamiltoniano) Percorso in un grafo che passa per tutti i nodi del grafo esattamente una volta.

PHAGE DISPLAY (esposizione sulla superficie di un fago) Tecnica in cui i fagi (batteriofagi) vengono ingegnerizzati per inserire peptidi estranei nelle proteine del loro capside in modo che questi vengano esposti sulla superficie del fago.

PHRED SCORE Punteggio della qualità della sequenza letta da una piattaforma di sequenziamento. È calcolato dalla probabilità che la lettura di un dato nucleotide sia errata, generalmente poi convertito in un numero intero rappresentato da un carattere ASCII e riportato in un file di formato FASTQ.

PIATTAFORMA Il computer o il sistema operativo su cui un certo programma è in grado di funzionare (o girare). Piattaforme diffuse sono: Windows, Macintosh, Unix e Linux.

PIRIMIDINA Composto azotato con una struttura ad anello. Timina e citosina sono basi pirimidiniche.

PIXEL (*PIcture ELement*) Elemento di immagine. Componente elementare di un'immagine (composta da numerosi pixel) costituito da un punto visualizzabile mediante illuminazione del display. Il pixel può variare per intensità, luminosità e colore. Più piccolo è il pixel, maggiormente definita e nitida risulta l'immagine.

PLASMIDE Elemento di DNA (in generale di origine batterica) con capacità di replicarsi indipendentemente dal cromosoma. I plasmidi vengono utilizzati per il clonaggio di sequenze estranee in cellule batteriche.

PLEIOTROPIA Effetto multiplo sul fenotipo di un organismo causato da un singolo gene o allele.

PLUG-IN Una porzione di software opzionale che aggiunge potenzialità a un software. I browser possono ospitare plug-in per mostrare file di diversi formati,

come per esempio file pdf o file di coordinate dal PDB (Capitolo 12), che possono essere visualizzati come molecole in 3D con l'apposito plug-in.

POLIMORFISMO Si definisce polimorfismo l'esistenza di due o più varianti alleliche di un gene, ciascuna delle quali assume proporzioni molto maggiori (generalmente almeno pari all'1% in una popolazione) di quelle attribuibili a mutazioni sporadiche o patologiche.

POLIPEPTIDE Singola catena di residui aminoacidici uniti da legami peptidici.

PONTI DISOLFURO Legami chimici che possono unire coppie di residui aminoacidici di cisteina. I ponti disolfuro in generale stabilizzano la struttura delle proteine che li contengono.

PRIMER Sequenza di innesco per dare inizio alla replicazione del DNA o a processi utilizzati in varie metodologie sperimentali (sequenziamento con il metodo di Sanger, PCR ecc.)

PROBE Vedere sonda.

PROCARIOTI Organismi che a differenza degli eucarioti sono privi di un nucleo circondato da membrane. Appartengono ai procarioti i batteri (o eubatteri) e gli archei (o archeobatteri).

PROCEDURA Insieme sequenziale delle azioni e dei percorsi per la soluzione di un problema.

PROCESSO MARKOVIANO Processo stocastico in cui lo stato futuro di un sistema dipende dallo stato corrente del sistema stesso.

PROFILO (o PSSM, *Position Specific Scoring Matrix*) Una tabella (matrice) derivata da un allineamento multiplo di proteine omologhe e contenente valori posizione-specifici e *gap penalty*. Ogni posizione di un profilo descrive una posizione dell'allineamento multiplo e consiste in una riga di valori (uno per ogni aminoacido) e in due valori relativi alla *gap penalty* e alla *gap extension penalty*. I valori contenuti nella riga sono ricavati dalle frequenze dei singoli residui nella posizione descritta e da una matrice di sostituzione (Capitolo 4).

PROGRAMMA Sinonimo di software. Sequenza di istruzioni rigorose e codificate, generate in un linguaggio di programmazione, che hanno lo scopo di comunicare al computer, in modo comprensibile, i compiti da svolgere.

PROGRAMMAZIONE Attività finalizzata a scrivere, mediante i linguaggi di programmazione, programmi eseguibili dal computer, allo scopo di svolgere operazioni utili o per risolvere problemi. Le fasi per la realizzazione di un programma sono: analisi del problema, stesura delle specifiche, definizione dell'algoritmo risolutore, codifica del programma mediante un linguaggio di programmazione, compilazione in linguaggio macchina, test.

PROGRAMMAZIONE DINAMICA Metodi algoritmici per la soluzione di problemi complessi, dividendoli in sotto-problemi di più facile soluzione. Esempi di algoritmi di programmazione dinamica usati in bioinformatica sono quelli di Needleman e Wunsch e di Smith e Waterman.

PROMOTORE Una regione di DNA, riconosciuta dall'enzima RNA polimerasi, necessaria per l'inizio della trascrizione.

PROTEINA Macromolecola composta da una catena di aminoacidi uniti da legami peptidici. Le proteine naturali sono composte da 20 aminoacidi codificati dal codice genetico (alanina, arginina, triptofano ecc.), più alcuni aminoacidi modificati, come selenocisteina e pirolisina.

PROTEOMA L'intero insieme delle proteine di un organismo.

PROTEOMICA Lo studio del proteoma.

PSEUDOGENE Sequenza genomica, generalmente derivante da un evento di duplicazione o retrotrascrizione, che ha accumulato mutazioni tali che la sequenza sia ancora riconoscibile come simile a un gene, ma non può più essere trascritta, oppure i suoi prodotti di trascrizione non sono funzionali.

PSSM (*Position Specific Scoring Matrix*) Vedere profilo.

PUNTEGGIO (*score*) Valore assegnato a una soluzione prodotta da un programma (per es. il punteggio assegnato a un allineamento tra sequenze viene calcolato come somma dei valori della matrice di sostituzione utilizzata dal programma e dalle penalità associate ai gap, vedere anche Capitolo 5).

PURINA Composto azotato con una struttura a doppio anello. Adenina e guanina sono basi puriniche.

P-VALUE Probabilità dell'ipotesi nulla in un test statistico. Valori piccoli di p-value possono portare a rigettare l'ipotesi nulla e implicitamente ad accettare l'ipotesi alternativa (vedere Capitolo 3).

PWM (*Position Weight Matrix*) Matrice calcolata partendo da una collezione di biosequenze accomunate da una caratteristica funzionale comune e allineate fra di loro. Ogni colonna della PWM descrive una diversa posizione dell'allineamento, mentre ogni riga riporta i caratteri che compongono l'alfabeto delle sequenze considerate. I valori nelle celle della matrice sono calcolati dalle frequenze dei caratteri dell'alfabeto in ogni posizione dell'allineamento.

QUERY L'input (sequenza o altro tipo di dato) da confrontare con tutte le entry di una banca dati.

RAM (*Random Access Memory*) Memoria elettronica primaria del computer caratterizzata da un'altissima velocità in lettura e scrittura, che contiene programmi, dati ed elaborazioni. La RAM è una memoria ad accesso casuale volatile perché, spegnendo il computer, si azzera. La sua capacità si misura in byte (megabyte, gigabyte).

READ Lettura della sequenza di un acido nucleico, o di una sua porzione, prodotta da una piattaforma di sequenziamento. È generalmente riportata in file di formato FASTQ, in cui sono contenuti anche i punteggi di qualità dei nucleotidi letti (PHRED score).

RECESSIVO Ogni carattere che viene espresso fenotipicamente solo quando presente in entrambi gli alleli di un genoma diploide.

REGEXP Vedere espressione regolare.

REPLICAZIONE La sintesi di una macromolecola identica a una data (per es. DNA).

REPRESSORE Prodotto di un gene regolatore che si lega a un sito specifico del DNA inibendo la trascrizione di uno o più geni.

RETE DI COMPUTER Sistema costituito da un insieme di computer, localizzati fisicamente in luoghi diversi e tra loro interconnessi per via telematica.

RETE NEURALE ARTIFICIALE (*artificial neural network*, ANN) Metodo di machine learning supervisionato basato sulla propagazione e modificazione di segnali in una rete di nodi a partire da un insieme di dati di *training* (allenamento).

RETRIEVAL Letteralmente recupero, termine utilizzato per indicare la selezione e l'estrazione di informazioni da una banca dati.

RICOMBINAZIONE Riarrangiamento di alleli dovuto a *crossing-over* o assortimento indipendente.

RNA Acido nucleico composto di ribosio, fosfato e di 4 nucleotidi: citosina, uracile, guanina e adenina. I principali tipi di RNA sono: l'RNA messaggero (mRNA), l'RNA transfer (tRNA) e l'RNA ribosomiale (rRNA).

RNA EDITING Serie di processi post-trascrizionali e co-trascrizionali che alterano la sequenza di un RNA modificandone i nucleotidi.

RNA MESSAGGERO (mRNA) La copia complementare di RNA, formata a partire da un singolo filamento di DNA attraverso il processo della trascrizione. Dopo la trascrizione, l'mRNA viene trasportato nel citoplasma e portato a maturazione (*splicing*) fino a diventare una sequenza contenente l'informazione necessaria a codificare una proteina.

RNA-SEQ Tecnica di sequenziamento parallelo massivo del trascrittoma di un campione, basato su retrotrascrizione in cDNA degli RNA, sequenziamento, mappatura/ricostruzione dei trascritti e quantificazione della loro espressione.

ROM (*Read Only Memory*) Memoria elettronica primaria a sola lettura che contiene programmi e dati immodificabili dall'utente. La sua capacità si misura in byte.

ROOT MEAN SQUARE DEVIATION (RMSD) Misura della distanza spaziale fra due serie di punti, in base a differenze delle coordinate cartesiane di punti equivalenti (Box 14.2).

ROUTER Computer dedicato alla connessione di una rete locale a Internet. I router controllano l'indirizzo di destinazione dei pacchetti che li attraversano e li indirizzano su uno specifico percorso (*route*).

SAM (*Sequence Alignment/Map*) Formato di interscambio per allineamenti di sequenze a un genoma di riferimento. La versione binaria di un file SAM è detta BAM. L'allineamento è descritto in forma compatta secondo la codifica CIGAR.

SCAFFOLD Struttura prodotta da algoritmi di assemblaggio di genomi, costituita da una serie di CONTIG ordinati e collocati reciprocamente a una distanza approssimata.

SCORE Vedere punteggio.

SEQUENZA SEGNALE Corta sequenza, in generale localizzata all'amino-terminale e capace di indirizzare una proteina all'interno del reticolo endoplasmatico. Esistono altri *sorting signal* (segnali di smistamento) responsabili dalla corretta locazione delle proteine nei

diversi compartimenti cellulari (mitocondri, cloroplasti, nucleo ecc.).

SEQUENZIAMENTO DEL DNA La tecnica con cui si decifra la sequenza delle basi di una regione di una particolare molecola di DNA o RNA.

SEQUENZIAMENTO PAIRED-END Tecnica di sequenziamento in cui si leggono entrambe le estremità dei frammenti di genoma sottoposti all'indagine. Solo alcune piattaforme di sequenziamento NGS, per es. quelle basate su tecnologia Illumina, sono in grado di effettuare questo tipo di lettura.

SEQUENZIAMENTO SINGLE END Tecnica di sequenziamento in cui si legge solo un'estremità dei frammenti di genoma sottoposti all'indagine.

SFF (*Standard Flowgram Format*) Formato di file binario, simile al formato FASTQ, prodotto in output dalle piattaforme di sequenziamento Roche 454.

SIMILARITÀ Misura che si può associare a un allineamento tra sequenze proteiche e che si ottiene sommando i valori associati nella matrice di sostituzione prescelta alle coppie di residui allineati (vedere identità).

SISTEMA OPERATIVO (OS, *Operating System*) Software di base, costituito da programmi e da routine, che consente, mediante semplici comandi logici, l'interazione utente-computer. Inoltre il sistema operativo controlla e supervisiona l'hardware rendendo il suo funzionamento ottimale. Risiede su disco e viene letto automaticamente all'accensione del computer in fase di inizializzazione. I più noti e diffusi sistemi operativi sono Windows, Macintosh, Unix e Linux.

SITO ATTIVO I residui aminoacidici componenti il sito catalitico di un enzima. I residui del sito attivo sono indispensabili alla funzione dell'enzima.

SITO DI SPLICING La sequenza che si trova al 5' e al 3' delle giunzioni esone/introne (vedere Capitolo 5).

SMITH E WATERMAN, ALGORITMO DI Algoritmo di programmazione dinamica per l'allineamento locale di due biosequenze.

SNP (*Single Nucleotide Polymorphism*) Variazioni di singoli nucleotidi nel genoma. Si pensa che ci siano svariati milioni di SNP nel genoma umano.

SOFTWARE Letteralmente oggetto soffice, in contrapposizione a hardware. Il termine indica tutto ciò che è relativo alla programmazione e cioè l'insieme dei programmi, dei documenti, delle regole e delle procedure che consentono al computer di risolvere problemi relativi all'elaborazione dei dati. Esistono varie tipologie di software: software di base (sistemi operativi per facilitare l'interazione uomo-computer), di utilità (per facilitare la gestione dell'hardware o del software), applicativi (per la soluzione di problemi specifici), didattici (per l'uso in ambienti educativi e formativi).

SONDA (*probe*) Ogni molecola che sia marcata in qualche modo (per es. con radioattivi o fluorescenti) in modo che possa essere usata per identificare o isolare un gene, un RNA o una proteina.

SPLICING Il processo al termine del quale una molecola di RNA trascritto da un gene contenente esoni e introni si trova composta con le sole sequenze esoniche.

SPLICING ALTERNATIVO Una delle combinazioni alternative di un RNA o di una proteina dovute al diverso possibile arrangiamento di segmenti di geni durante lo splicing dell'RNA messaggero che avviene negli organismi eucarioti.

STRUTTURA SECONDARIA Nelle proteine, indica l'organizzazione del *backbone* proteico in strutture quali l'α-elica e il filamento o il foglietto β (vedere Capitolo 14). Negli RNA indica il pattern di legami idrogeno Watson-Crick intramolecolari (vedere Capitolo 11).

STRUTTURA TERZIARIA Avvolgimento di una catena polipeptidica o di un acido nucleico in una struttura tridimensionale attraverso la formazione di legami idrogeno, idrofobici e, nelle proteine, ponti disolfuro.

SUBDIRECTORY Directory presente all'interno di un'altra directory. Va notato che tutte le directory sono subdirectory della directory principale (*Root directory*) di un disco.

TCP/IP (*Transmission Control Protocol* – TCP – e *Internet Protocol* – IP) Protocolli su cui si basa il funzionamento logico della rete Internet.

TGS (*Targeted Genome Sequencing*) Sequenziamento di una porzione del genoma, selezionata in base allo scopo dell'analisi (per es. il sequenziamento solo dei geni noti per essere potenzialmente responsabili di una patologia).

THREADING Classe di algoritmi per la predizione della conformazione che una sequenza proteica può assumere.

TRANSIZIONE Mutazione che porta alla sostituzione di una purina con un'altra purina, o di una pirimidina con un'altra pirimidina.

© 978-88-08-**62112**-2

TRANSVERSIONE Mutazione che porta alla sostituzione di una purina con una pirimidina, o viceversa.

TRASCRITTOMA L'insieme degli RNA che possono essere trascritti a partire da un genoma.

TRIMMING Procedura di pulizia di un insieme di read prodotte da sequenziamento NGS. Queste procedure possono rimuovere porzioni di read i cui punteggi PHRED di qualità sono insufficienti, oppure scartare intere read di bassa qualità media o rimuovere sequenze di adattatori, e altro ancora.

UNITÀ TASSONOMICHE OPERATIVE (OTU) Definizione operativa di un gruppo di individui. Può coincidere con la specie.

VARIANT CALLING Procedura per l'identificazione di varianti dal sequenziamento del genoma estratto dalle cellule di un campione, rispetto a una sequenza genomica usata come riferimento. Il risultato di tali procedure è generalmente riportato nel formato di interscambio VCF.

VCF Formato di file di interscambio per riportare variazioni genetiche in un campione rispetto alla sequenza di un genoma usato come riferimento, prodotto da software di variant calling.

WES (*Whole Exome Sequencing*) Sequenziamento delle porzioni di un genoma corrispondente solo agli esoni.

WGS (*Whole Genome Shotgun*) Sequenziamento di un intero genoma, partendo dalla frammentazione casuale di tutto il DNA estratto dalle cellule di un campione.

YAC (*Yeast Artificial Chromosome*) Vettore usato per il clonaggio di porzioni estese di DNA esogeno, basato sulla struttura e sull'organizzazione dei cromosomi di lievito.