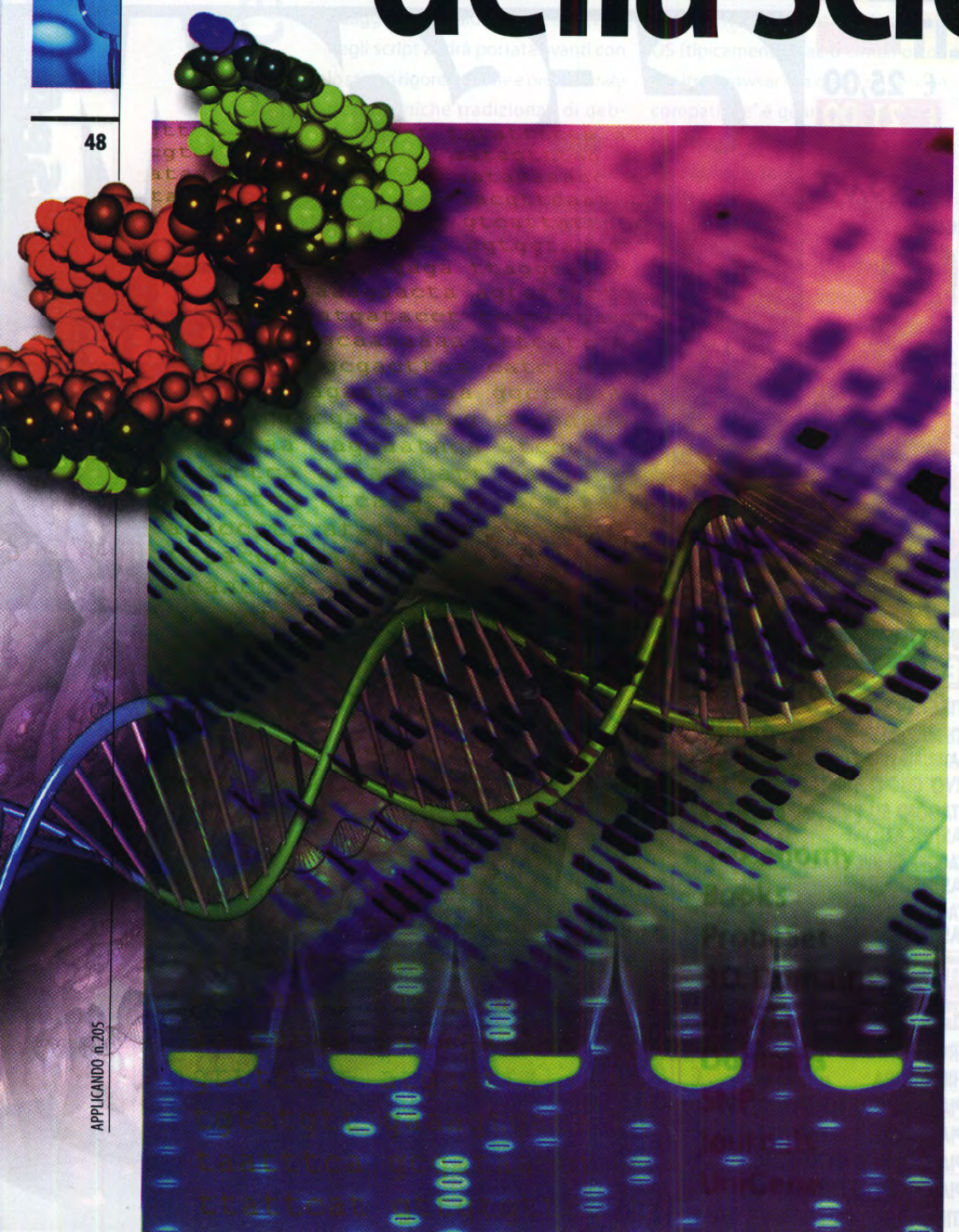
Viaggio nei meandri dei software e dei siti indispensabili per la ricerca scientifica e per l'elaborazione dei dati

al servizio della scienza

di Beniamino Cenci Goga ▶ beniamino.cencigoga@applicando.com



Non solo statistica, non solo grafici e non solo scienza. La vita professionale del ricercatore, sia esso un borsista alle prime armi o un affermato scienziato, è ormai un compromesso tra studio, pianificazione degli esperimenti, esecuzione della sperimentazione e quindi di acquisizione, elaborazione e presentazione dei dati. Se a questo aggiungiamo le innumerevoli ore di insegnamento e il tempo dedicato al reperimento e, troppo spesso, all'amministrazione dei fondi per le attività di ricerca, ci rendiamo facilmente conto di quanto sia necessario avere piena padronanza dei software scientifici e per l'elaborazione dei dati. I nostri colleghi d'oltre oceano possono contare su specifici curriculum dedicati alla statistica e all'epidemiologia. Esistono anche specifiche figure professionali, il cui compito è quello di elaborare le richieste di finanziamento e portare a termine la loro gestione amministrativa. Nel nostro Paese invece, salvo rare eccezioni, il ricercatore volonteroso deve impegnarsi su più fronti, spesso subendo anche le angherie da parte di superiori meno motivati! In questo confuso panorama, l'unica via di uscita appare l'ottimizzazione

▼ figura 1

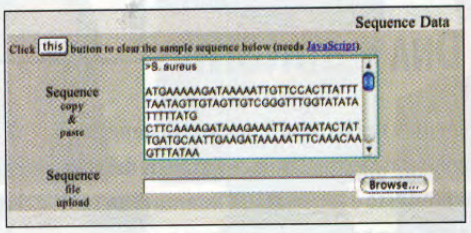
```
ORIGIN
        1 atgaaaaaga taaaaattgt tccacttatt tta
       61 tatttttatg cttcaaaaga taaagaaatt aat
      121 aatttcaaac aagtttataa agatagcagt tat
      181 gaaatgactg aacgtccgat aaaaatatat aat
      241 caggatcgta aaataaaaaa agtatctaaa aat
      301 attaaaacaa actacggtaa cattgatcgc aac
      361 ggtatgtgga agttagattg ggatcatagc gtc
      421 agcatacata ttgaaaattt aaaatcagaa cgt
      481 gaattggcca atacaggaac acatatgaga tta
      541 aaagattata aagcaatcgc taaagaacta agt
      601 tggatcaaaa ttgggtacaa gatgatacct tcg
      661 gatgaatatt taagtgattt cgcaaaaaaa ttt
```

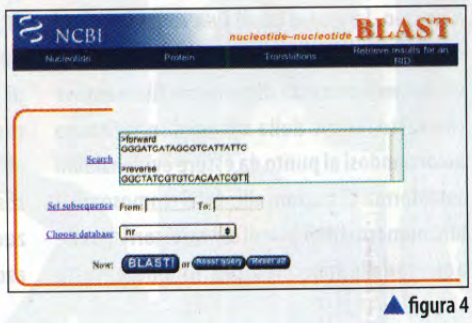
del tempo, almeno fino a quando la nostra classe politica non vorrà porre rimedio con precisi provvedimenti alla fuga e alla compressione dei cervelli.
L'ottimizzazione del tempo è un argomento che sta particolarmente a cuore ai Mac User, se è vero che, grazie alla linearità delle applicazioni per Mac OS, l'utente Mac risparmia diverse decine di ore all'anno nei confronti degli utenti di altri Sistemi Operativi.
Il ricercatore lungimirante ha così imparato a proprie spese che, se è vero che "chi fa da sé fa per tre", è ancor più vero che un ottimo supporto informatico e un bagaglio culturale di bit e byte potranno colmare il divario che ci allontana dai colleghi americani.
Nella nostra trattazione ci soffermeremo sui più recenti software per la biologia molecolare, quindi sulle applicazioni per l'elaborazione dei dati e infine forniremo i suggerimenti per la scelta e l'utilizzo delle migliori applicazioni per la presentazione dei risultati.

Il DNA a portata di mouse

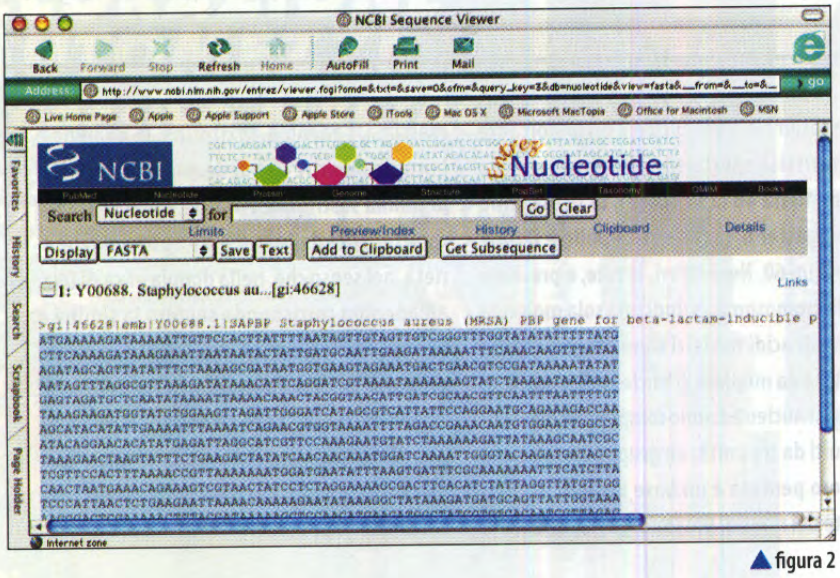
Il premio Nobel per la chimica Kary Mullis nel 1983 scoprì un processo che avrebbe dato nuovo impulso alla ricerca. Nel 1985 Mullis, che tra l'altro è un Mac User, descrisse nella rivista *Scientific American* la sua scoperta, derivata da una visione in una notte di luna piena a bordo della sua auto nelle montagne della California, della PCR (Polymerase Chain Reaction), cioè di una reazione che consente di produrre copie esatte e in maniera selettiva ed esponenziale di una determinata molecola di DNA. Le applicazioni sono tantissime: diagnostica forense, diagnostica di malattie infettive, studi di evoluzione, identificazione di malattie genetiche, e altro ancora. La reazione si basa sulla capacità di due oligonucleotidi, detti primer (cioè dei nucleotidi costituiti da poche basi), con sequenze di basi azotate complementari alla molecola su cui si trova il tratto di cui fare copie multiple (amplificazione) (vedi box: DNA e dintorni).
Prima dell'avvento dei software per personal computer o residenti su siti web, al ricercatore non restava che armarsi di pazienza e "disegnare" la coppia di oligonucleotidi con procedure non proprio alla portata di tutti i ricercatori del tempo, partendo dalla sequenza di basi descritta da altri (**figura 1**) e andando per tentativi. Ora non è più così e sono disponibili diverse possibilità, sia gratuite online, sia sottoforma di pacchetti software. La scelta autarchica inizia con una visita al sito di PubMed (www.ncbi.nlm.nih.gov), dove nella specifica sezione è possibile andare alla ricerca del gene che ci interessa, sia digitandone il nome, sia il numero di accesso, se noto. Il software residente nel sito si incarica di andare alla ricerca del gene nella banca dati e in pochi secondi ce lo mostra nella sua interezza (**figura 2**). A questo punto non resta che passare a un sito nel quale un altro software residente consentirà di selezionare la coppia di oligonucleotidi necessaria per la copia multipla di parte del gene. Un ottimo sito ad accesso ancora gratuito è http://bibiserv.techfak.uni-bielefeld.de/genefisher/. Da qui è possibile inserire la sequenza ottenuta in precedenza in

▼ figura 3





▲ figura 4



▲ figura 2

figura 5

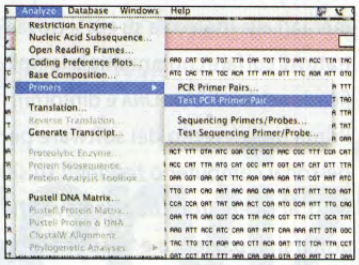


figura 7

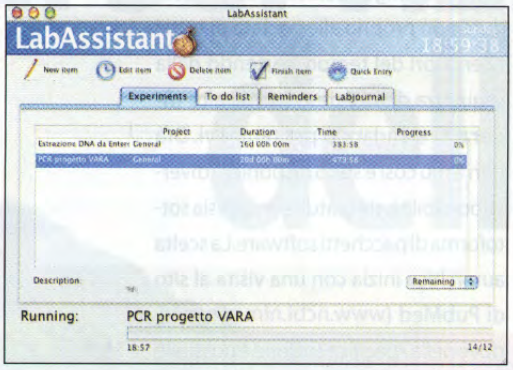
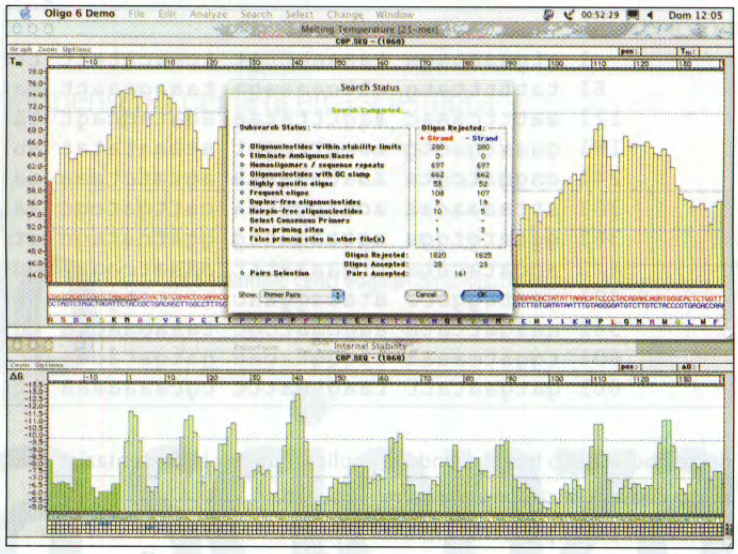


figura 6

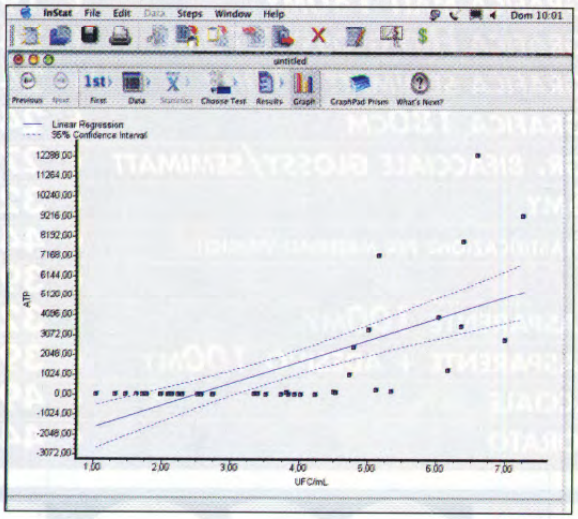


un apposito campo (**figura 3**) e impostare le nostre condizioni sperimentali: in breve ci verrà proposta una lista di coppie di oligonucleotidi e le loro caratteristiche. La specificità della coppia eventualmente selezionata può essere condotta sul sito http://bioweb.pasteur.fr/seqanal/interfaces/dialign2-simple.html dove, in un apposito campo, è possibile inserire la sequenza del gene e quelle dei due oligonucleotidi, in attesa del corretto allineamento. I puristi non si accontentano e, attraverso una specifica sezione del sito di PubMed detta BLAST (**figura 4**) vanno alla ricerca di tutte le possibili corrispondenze presenti nelle banche dati. Spesso è utile scoprire che la coppia di oligonucleotidi selezionata è in grado di effettuare copie multiple (di amplificare) anche di tratti di DNA appartenenti a specie completamente differenti: se le dimensioni del prodotto amplificato sono simili il rischio di analisi falsate è in agguato. Di tempo rispetto ai pionieri della biologia molecolare ne abbiamo risparmiato, ma il danzare da un sito all'altro, soprattutto se le sequenze da analizzare sono molte, può essere fonte di errori e confusione. Esistono, in effetti, degli ottimi software che, a prezzi non proprio da saldo e comunque non alla portata dell'utente privato, permettono di fare tutto o quasi in remoto, limitando la connessione a Internet alla consultazione delle banche dati.

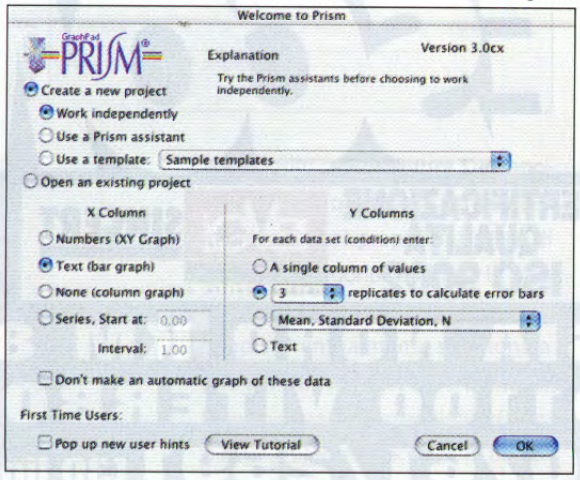
Per Mac OS X sono disponibili due ottimi prodotti: MacVector di accelrys e Oligo di Molecular Biology Insight. La ricerca del gene in MacVector è precisa e la selezione della coppia di oligonucleotidi (primers per gli anglofoni) rapidissima, attraverso il menu Analyze>Primers> PCR Primer Pairs. La verifica dell'allineamento si ottiene attraverso il comando Test PCR Primer Pair... (**figura 5**). L'utilizzo di Oligo è altrettanto agevole, anche se le funzioni sono minori di quelle offerte da MacVector. Non mancano una sezione per la ricerca di primer e sonde (**figura 6**) e la visualizzazione grafica dell'allineamento degli oligonucleotidi. Chi dedica le sue energie allo studio degli enzimi troverà un valido aiuto nell'ottimo freeware EnzymeX. Ora tutto è pronto per l'acquisto dei reagenti e

DNA e dintorni

Mentre nei batteri il materiale genetico è mescolato al citoplasma, nelle cellule dei mammiferi esso è segregato all'interno di un territorio, circondato da un involucro membranoso, il nucleo. La quantità di DNA contenuto nel nucleo è costante per ogni specie animale. Vi è un numero costante di molecole filamentose che verso l'inizio della mitosi si spiralizzano accorciandosi al punto da essere evidenziabili sottoforma di bastoncelli, detti cromosomi. Il loro numero si può quindi contare: corrisponde a quello delle molecole di DNA. Il numero esatto di cromosomi in ogni nucleo è un'importante caratteristica specie-specifica: ad esempio l'uomo ne ha 46 (23 coppie), la Drosophila melanogaster 8, la rana 26, il pollo e il cane 78, i bovini 60. Nei batteri, invece, è presente un solo cromosoma, quindi una sola molecola di DNA. Gli acidi nucleici sono lunghe catene costituite da migliaia di nucleotidi legati uno all'altro. I nucleotidi sono complessi molecolari costituiti da tre unità: un gruppo fosfato, uno zucchero pentoso e un base azotata. Le basi azotate dei ribotidi possono essere, in linea di massima, l'adenina, la citosina, la guanina e l'uracile, mentre in quelle dei deossiribotidi la timina sostituisce l'uracile. Un'importante caratteristica delle basi è la loro complemetarietà, nel senso che, nella doppia elica di DNA, all'adenina corrisponde sempre la timina e alla citosina la guanina (A-T; C-G). La struttura primaria del DNA è la specifica sequenza dei singoli nucleotidi nelle catene polinucleotidiche. Si tratta di una informazione importantissima in quanto alla base della conoscenza del codice genetico.



▲ figura 8



▲ figura 9

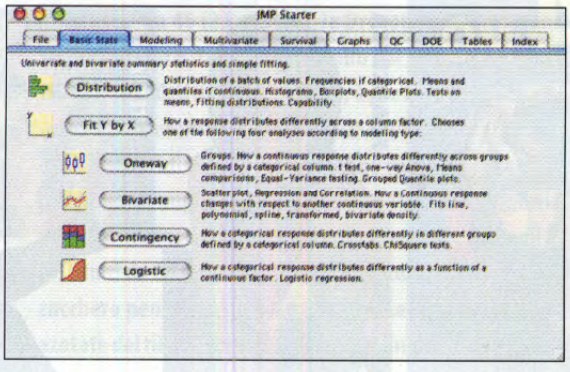
per l'inizio della sperimentazione: LabAssistant è un freeware che può essere d'aiuto a chi proprio non riesce a farsi assegnare dei validi collaboratori (**figura 7**).

Statistica e piani sperimentali

Se tutto è andato per il verso giusto e il nostro entusiasmo non è stato tarpato, siamo pronti a pianificare il nostro esperimento. Con il termine pianificazione degli esperimenti si intende non tanto la formulazione degli scopi o della suddivisione di una ricerca in pacchetti (i cosiddetti work-package degli anglofoni), quanto quell'insieme di accorgimenti tecnici, presi all'atto della stesura del protocollo sperimentale, necessari per ottenere risultati validi dal punto di vista statistico. Un esperimento ben pianificato è basato su domande formulate in modo preciso e prevede che le procedure sperimentali rispondano in maniera inequivocabile ai quesiti. Prevalentemente il ricercatore si trova di fronte ai cosiddetti esperimenti comparativi, il cui scopo risiede nel confronto degli effetti di determinati trattamenti sulle unità sperimentali. Di altro tipo sono le indagini statistiche nelle quali il ricercatore ha un ruolo passivo, non potendo intervenire sui fenomeni dei quali si desiderano conoscere le caratteristiche. Diremo, per inciso, che troppo spesso in campo scientifico, al solo scopo di produrre dei dati per assecondare la terribile legge "publish or perish" del mondo accademico scientifico, ci si limita a osservazioni statistiche o epidemiologiche basate su un numero limitato di casi, dimenticando l'approccio di curiosità tipico delle persone ingegnose, di quelle che "riescono a cambiare le cose, che fanno progredire l'umanità" come diceva Dario Fo nella versione italiana della famosa campagna pubblicitaria Apple, Think Different ("they push the human race forward" nella versione in lingua originale).

Anche se non apparteniamo alla categoria di "coloro che essendo abbastanza folli da pensare di cambiare il mondo lo cambiano davvero" (Dario Fo, Think Different, Apple), dopo l'impostazione e l'esecuzione dell'esperimento si impone una scelta obbligata: la scelta e l'utilizzo di un software per l'elaborazione dei dati raccolti. Al neofita suggeriamo l'ottimo InStat di GraphPad Software. InStat non è stato sviluppato da matematici, ma da ricercatori e l'approccio, anche se sconcertante per i duri e puri utenti dei software professionali prodotti da SAS Institute, è estremamente lineare. La prima schermata consente infatti di definire gli obiettivi e il tipo di dati di cui disponiamo, quindi sarà sempre il software a guidarci passo-passo verso le scelta dei test da effettuare, la loro visualizzazione e persino la rappresentazione grafica (**figura 8**). Ai più raffinati, desiderosi di un output elegante e personalizzato, è destinato Prism della stessa sowtware house, che nella splendida versione per Mac OS X (**figura 9**) rende l'elaborazione statistica un'incombenza davvero piacevole. Anche in questo caso la guida passo-passo accompagna l'utente meno smaliziato sino alla fine dell'elaborazione, consentendogli di dedicarsi a infinite personalizzazioni. Non mancano la possibilità di inserire agevolmente caratteri di altri alfabeti e di esportare le nostre fatiche in formati non proprietari, per la successiva elaborazione con altri software. La quasi totalità dei comandi è accessibile dalla barra dei menu, pregevole esempio di razionalità e funzionalità che vorremmo vedere imitato anche da software house più famose. Prism offre anche la peculiarità di visualizzare dati con rappresentazione grafica degli stessi nella medesima schermata. Una soluzione più professionale, ma che condivide con InStat l'attuale tendenza degli sviluppatori di pacchetti per la statistica, è JMP prodotto da una consociata di SAS Institute. Anche JMP, come InStat, permette di selezionare quali test saranno effettuati e fornisce un valido aiuto nell'inserimento dei dati (**figura 10**). Le possibilità offerte sono superiori a quelle di InStat, al costo di una leggera complicazione nella scelta dell'output. Non mancano nemmeno a questo pacchetto le possibilità di intervento manuale per modificare l'aspetto dei grafici. È ovvio che gli utenti classici

▼ figura 10



dei pacchetti professionali di SAS, tra i quali ricordiamo l'ottimo StatView, purtroppo fermo alla versione per Mac OS 9, considerano un'inutile complicazione il cosiddetto inserimento facilitato dei dati, ma i tempi stanno cambiando e la statistica non è più esclusivo appannaggio dei matematici. Ricordiamo, a uso dei neofiti, che i pacchetti classici prevedono l'inserimento dei dati in colonna, contraddistinti da un'etichetta univoca, mentre le applicazioni più moderne, come quelle di cui ci stiamo occupando, permettono di optare per inserimenti più intuitivi, con conseguente risparmio di tempo. È però vero anche il rovescio della medaglia: in effetti, dopo anni di utilizzo dei software classici, la nostra mente si era arresa all'apparente illogicità, e non nascondiamo che sulle prime InStat e JMP ci hanno leggermente disorientato. Sono però bastate poche ore di utilizzo per convincerci della bontà dell'approccio semplificato.

La scelta autarchica

Microsoft Excel, soprattutto nella versione X, è un prodotto maturo e può rappresentare un valido aiuto per chi non volesse investire in un pacchetto specifico. In effetti Excel è dotato di moltissimi algoritmi per la statistica, cioè di formule che consentono di giungere alle stesse conclusioni offerte dai software descritti in precedenza, al prezzo di un po' di clic supplementari. L'output, soprattutto grazie al layer di disegno antialias Quartz, è interessante e in molte situazioni non fa rimpiangere scelte più costose (**figure 11 e 12**). Tuttavia, se la presentazione dei nostri risultati statistici merita di più, perché non provare il neonato ConceptDraw Pro di Computer Systems Odessa (**figura 13**)? ConceptDraw Pro è una piacevole applicazione di disegno che permette, a un prezzo interessante, di intervenire sui nostri grafici attraverso le molteplici funzioni di cui dispone (**figura 14**). La nostra esperienza ormai decennale nel campo della biologia molecolare e della statistica ci ha permesso di passare in rassegna software e siti dall'interfaccia amichevole, ancorché di estremo aiuto per il ricercatore moderno, nella consapevolezza di esporre argomenti seri con linguaggio leggero. Tuttavia, e questa raccomandazione è ricordata spesso in tutti i manuali d'uso delle applicazioni descritte nell'articolo, l'utilizzo di software e siti specifici non può prescindere dall'approfondita conoscenza delle basi scientifiche e matematiche. Vogliamo comunque credere che l'associazione tra software ben scritti, siti web gradevoli e Macintosh possa rappresentare uno stimolo in più per approfondire le nostre conoscenze.

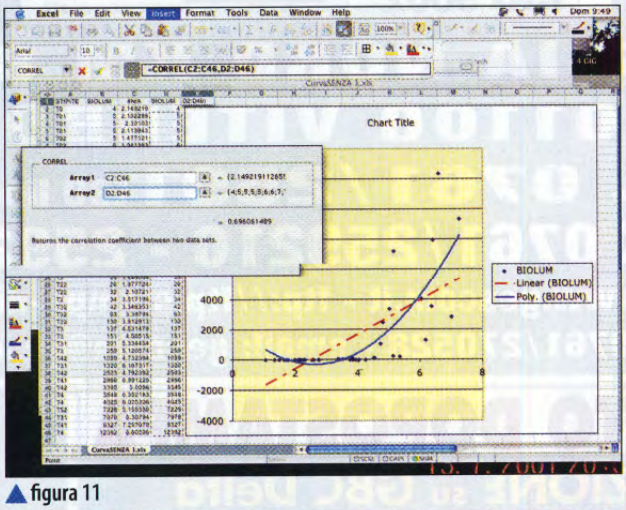


figura 11

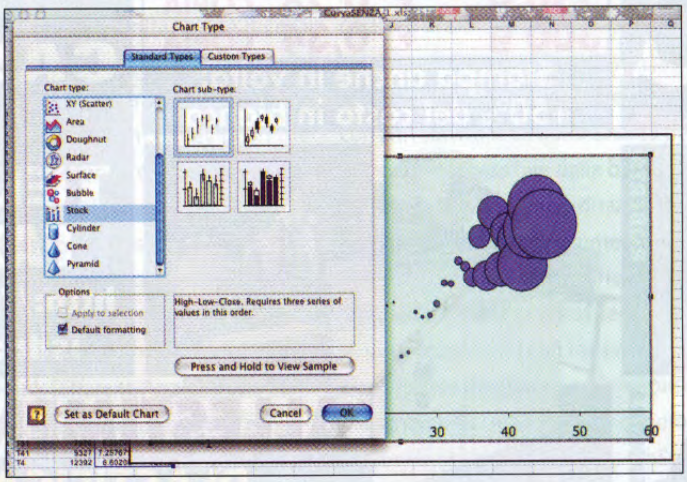


figura 12

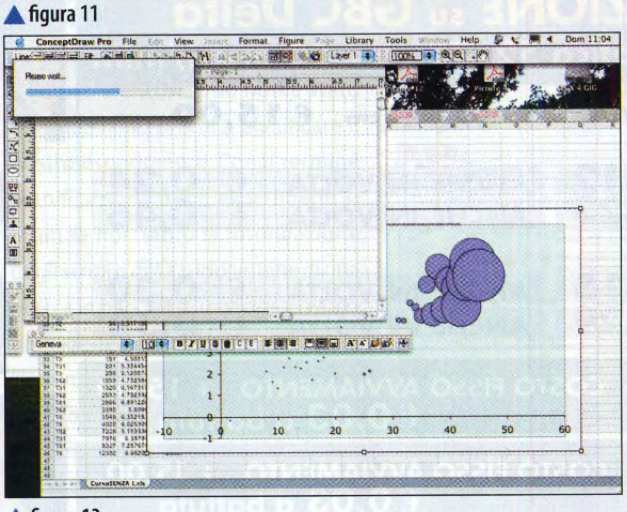


figura 13

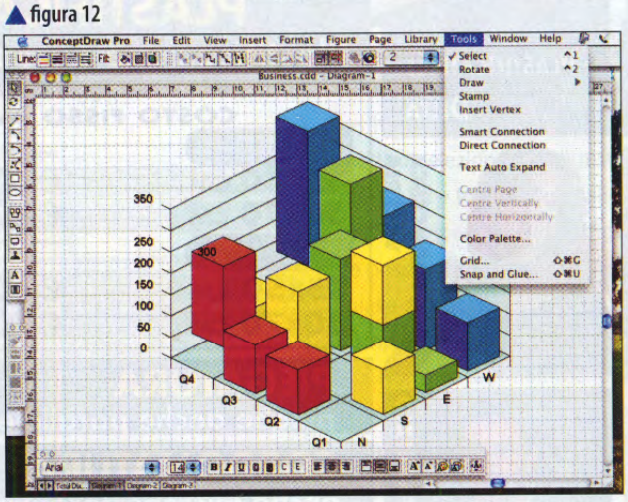


figura 14