# PHẦN DẪN NHẬP

# NGÀNH SINH TIN HỌC

## 1.Giới thiệu

Dữ liệu sinh học đang được thu nhận với tốc độ vũ bảo. Đến tháng 8 năm 2000, ngân hàng dữ liệu GENEBANK đã có 8,214,000 mục liên quan đến các trình tự sinh học DNA [2] và cơ sở dữ liệu (CSDL) SWISS-PROT có 88,166 mục liên quan đến các trình tự protein[3]. Trung bình những CSDL đang tăng gấp đôi kích thước sau mỗi chu kỳ 15 tháng [2]. Thêm vào đó, việc công bố bộ gen của hơn 40 bộ phận cơ thể đã cung cấp thêm từ 450 gen đến trên 100,000 gen. Ngoài ra có vô số dự án nghiên cứu gen, xác định cấu trúc protein được mã hóa trong bộ gen... đã sản sinh một lượng lớn thông tin sinh học và thông tin này ngày càng đa dạng và phong phú.

## 2.Đinh nghĩa sinh tin học

Do dữ liệu sinh học tăng trưởng mạnh mẽ nên công cụ tin học đã trở thành một phương tiện không thể thiếu trong phân tích xử lý dữ liệu sinh học. Công nghệ Thông tin có thể quản lý nguồn dữ liệu khổng lồ, phân tích các dữ liệu đa dạng và luôn biến đổi trong thế giới tự nhiên. Ngành Sinh tin học được xem là lĩnh vực nghiên cứu liên ngành nhằm kết hợp các kỹ thuật xử lý, tính toán và tổ chức thông tin bằng thiết bị Tin học với các kỹ thuật, công cụ phổ biến trong ngành sinh học phân tử.

Sự hợp nhất ngoài mong đợi giữa hai ngành khoa học thúc dẩy các nghiên cứu mạnh mẽ về công nghệ Sinh học đặc biệt là các nghiên cứu sinh lý học của một cơ phận ở mức độ gen.

Các tiến bộ nhanh chóng của kỹ thuật máy tính trong thu nhận dữ liệu cho phép dễ dàng thu nhận dữ liệu trình tự sinh học. Anthony Kerlavage của Công ty Celeron ( Mỹ) cho biết có thể dễ dàng tạo ra trên 100 GB dữ liệu trong một ngày [5].

### 3. Mục tiêu của Ngành Sinh Tin học:

Mục tiêu đầu tiên của ngành Sinh tin học là tổ chức dữ liệu để quản lý và truy cập thông tin. Mục tiêu thứ hai là phát triển các công cụ và tài nguyên hỗ trợ phân tích dữ liệu sinh học, chẳng hạn so sánh trình tự protein đặc thù với các trình tự đã biết rõ chức năng. Mục tiêu thứ ba là dùng những công cụ này để phân tích dữ liệu và diễn giải kết quả theo ý nghĩa trong sinh học.

Những nghiên cứu sinh học truyền thống thường kiểm tra hệ thống cá thể bằng cách so sánh chúng với các cá thể liên quan. Trong Ngành Sinh tin học, có thể quản lý những dữ liệu sinh học đã phân tích trên phạm vi toàn cầu thông qua mạng Internet và hỗ trợ tích cực các quá trình so sánh.

Phần tổng quan này tập trung vào mục tiêu thứ nhất và thứ ba. Đặc biệt, phần này sẽ bàn đến các nguồn dữ liệu hiện có, cách thức truy cập, phân tích . xử lý dữ liệu một số những ứng dụng thực tiễn của ngành Sinh tin học.

### 4. Thông tin kết hợp với sinh học phân tử.

Hãy bắt đầu bằng một cái nhìn khái quát về các nguồn tài nguyên thông tin. Chúng được chia thành các trình tự sinh học DNA, các trình tự sinh học protein, cấu trúc của các đại phân tử, trình tự gen và bộ gen khác.

Trình tự sinh học DNA thô là chuỗi được cấu tạo từ 4 ký tự cơ bản(nucleotide) là A, C, G, T. Trung bình mỗi gen có chiều dài khoảng 1,000 ký tự cơ sở ( base). Ngân hàng dữ liệu GENBANK hiện lưu trữ hơn 9.5 tỉ

nucleotide của các gen. Kế đến là các trình tự protein được cấu tạo từ 20 ký tự acid amino. Hiện có khoảng 300,000 trình tự protein đã biết, một protein của vi khuẩn có chiều dài 300 codon.

Dữ liệu cấu trúc đại phân tử là một dạng phức tạp của thông tin. Hiện có trong ngân hàng dữ liệu protein (PDB) có hơn 13,000 mục trình bày các cấu trúc protein.

Củng như những trình tự DNA thô, bộ gen bao gồm những ký tự cơ bản, có phạm vi từ 1.6 triệu đến 3 tỉ ký tự cơ sở. Điều quan trọng nhất của bộ gen hoàn chỉnh là khả năng phân biệt giữa vùng mã hóa và vùng không mã hóa. Giờ đây, có thể đo mức độ biểu hiện của hầu hết các gen trong từng tế bào trên toàn bộ gen. Những đo đạc mức độ biểu hiện được thực hiện trong những điều kiện môi trường, phạm vi hoạt động của chu kỳ tế bào, kiểu tế bào khác nhau trong hệ thống đa bào. Tập dữ liệu lớn nhất hiện có tương đương với số liệu 20 lần đo đạc cho 6,000 gen [10]. Dữ liệu thu được bao gồm các thông tin hoá sinh trong quá trình trao đổi chất, điều tiết, tương tác giữa protein-protein. . . .

Tính đa dạng và sự phức tạp của các tập dữ liệu khác nhau là vấn đề luôn tồn tại. Vẫn luôn luôn có nhiều dữ liệu trình tự thô hơn là dữ liệu có cấu trúc. Do đó đòi hỏi các khả năng phân tích số lượng khoẳng lồ các dữ liệu thô để thu nhận các thông tin có tính khái quát cao.

## 5. Tổ chức thông tin trên một diện rộng

Khái niệm cơ sở cho hầu hết các phương pháp nghiên cứu trong sinh tin học là có thể gom nhóm dữ liệu theo mức độ tương đồng và có ý nghĩa trong sinh học. Ví dụ, các đoạn trình tự sinh học thường được lặp lại tại những vị trí khác nhau của hệ gen DNA[11]. Gen có thể được gom thành các cụm có chức năng riêng biệt (ví dụ hoạt đông enzym) hay theo cách trao đổi chất của

chúng[13]. Ngoài ra có thể so sánh các protein chưa biết chức năng với các protein đã biết rõ chức năng để suy diễn chức năng và tiến hoá. Ở mức độ cấu trúc, hiện nay có một số hữu hạn các cấu trúc cấp ba khác nhau ( khoảng từ 1,000 đến 10,000 ) [14,15] và các protein có thể có cấu trúc tương đương nhưng khác nhau về trình tự .

Thuật ngữ chung để mô tả mối liên hệ giữa cặp protein hay gen với protein hay gen dùng để suy dẫn ra chúng: Các protein tương tự (analogous) có các nếp gấp có liên liên quan với nhau nhưng các trình tự thì không liên quan nhau. Trong khi các protein tương đồng (homology) giống nhau về trình tự và cấu trúc. Đôi khi rất khó phân biệt hai loại này đặc biệt nếu mối liên hệ giữa hai protein là xa nhau [17,18]. Trong quan hệ tương đồng cần phân biệt giữa orthologues - protein trong những loài được tiến hoá từ một gen tổ tiên chung, và paralogues - protein liên quan đến việc nhân đôi gen bên trong bộ gen [19] . Orthologue thường giữ lại chức năng giống nhau trong khi paralogue tiến hoá khác nhau nhưng có các chức năng có liên quan với nhau [20].

Khái niệm quan trọng nảy sinh từ quan sát này là ở chỗ các sinh vật khác nhau có “danh sách thành phần” hữu hạn [21,21]. Các protein trong một sinh vật được sắp xếp theo những thuộc tính khác nhau như trình tự gen, các nếp gấp protein hay chức năng của chúng. Ví dụ cấu trúc cấp ba của protein chỉ thích ứng với một số giới hạn các nếp gấp trong kho lưu trữ. Vì số các họ nếp gấp khác nhau là khá nhỏ so với họ gen, việc phân loại protein theo các nếp gấp làm đơn giản hóa trọng thông tin ẩn chứa trong bộ gen. Có thể cung cấp sự đơn giản hoá tương tự dựa trên các thuộc tính khác nhau như chức năng của protein.Do vậy, chúng ta rất mong danh sách các phần hữu hạn sẽ ngày càng phổ biến trong phân tích hệ gen.

Rõ ràng, vấn đề then chốt của việc quản lý lượng lớn dữ liệu này nằm ở nhu cầu phát triển các phương pháp truy vấn tương tự giữa các phân tử sinh học khác nhau và nhận diện những thứ có liên quan nhau. Phần sau bàn đến các CSDL chính cho phép truy cập tài nguyên thông tin và giới thiệu vài CSDL thứ cấp có gom nhóm dữ liệu. Các so sánh dễ dàng giữa bộ gen và sản phẩm của nó, cho phép nhận diện mối liên hệ và rút ra các đặc trưng nổi bật và duy nhất.

### 4.1. CSDL trình tự protein

CSDL trình tự protein được phân loại như sơ cấp, hổn hợp, thứ cấp. CSDL sơ cấp chứa trên 300,000 trình tự protein và chức năng của nó, đây là một kho lưu trữ dữ liệu thô. Một số kho lữu trữ dữ liệu phổ biến chung như SWISS-PROT, và PIR chứa các trình tự, chức năng protein, cấu trúc và những thay đổi sau khi dịch mã. CSDL hổn hợp như OWL [24] và NRDB [25] biên soạn và lọc dữ liệu trình tự các CSDL sơ cấp để tạo tập dữ liệu tổng hợp và hoàn chỉnh hơn dữ liệu thô của các CSDL riêng lẻ. CSDL này cũng bao gồm dữ liệu trình tự protein từ việc dịch mã các vùng mã hoá trong trình tự DNA.. CSDL thứ cấp gồm thông tin được suy dẫn từ các trình tự protein và giúp người dùng xác định trình tự mới có thuộc họ protein đã biết hay không. Một trong những CSDL phổ biến là PROSITE [26], đây là một CSDL chứa các mẫu trình tự ngắn và hồ sơ tổng lược (profile) nhằm biểu thị các vị trí (site) có ý nghĩa sinh học trong protein. CSDL PRINTS [27] mở rộng khái niệm này và cung cấp bản tóm tắt dấu vân protein – nhóm của đoạn lặp được bảo tồn để đặc trưng cho họ protein. Đoạn lặp thường cách nhau trong trình tự protein, nhưng vẫn liên tục trong không gian 3D khi protein bị gấp nếp. Bằng việc sử dụng nhiều đoạn lặp, có thể mã hoá các nếp gấp protein, các chức năng trong PROSITE . Cuối cùng, CSDL Pfam [28] chứa các phương án chỉnh thẳng cột nhiều trình tự (multiple alignment) và hồ sờ tổng

lược Markov ẩn của nhiều protein phổ biến. CSDL Pfam-A chứa các phương án chính thẳng cột trong khi Pfam-B là kết quả gom cụm tự động của toàn bộ dữ liệu CSDL SWISS-POT. Những CSDL thứ cấp khác biệt này được kết hợp lại với nhau thành tài nguyên duy nhất có tên là InterPro [29].

### 4.2. CSDL có cấu trúc

Kế đến hãy xem CSDL của cấu trúc đại phân tử. Ngân hàng dữ liệu protein, PDB [6,7], cung cấp tất cả cấu trúc 3D của các đại phân tử như protein, RNA, DNA và những phức hợp khác, hiên CSDL có chừng 13,000 cấu trúc (tháng 8 năm 2000) được phân giải bằng tia x và NMRb, ngoài ra còn có vài mô hình lý thuyết. PDBsum[30] cung cấp một trang Web riêng cho từng cấu trúc trong PDB chứa các chi tiết các phân tích cấu trúc, các biểu đồ và dữ liệu tương tác giữa các phân tử khác nhau.. Ba CSDL chính phân loại proteins theo cấu trúc để nhận diện các quan hệ về cấu trúc và tiến hóa là CSDL CATH[3], SCOP[32], và FSSP [33]. Các CSDL trên đều có kiến trúc phân cấp phục vụ việc gom nhóm protein dựa trên mức độ tương tự. CSDL khổng lồ này bao gồm những đại phân tử đặc biệt. Các CSDL này bao gồm CSDL Nucelic Acids, NDB[34] với các cấu trúc liên quan đến nucleic acids, CSDL HIV protease [35] chứa các cấu trúc protease HIV-1, HIV-2 và SIV và những phức hợp của chúng, và ReLIBase [36] chứa các phức hợ p receptor-ligand.

### 4.3. Trình tự nucleotide và gen

Như đã mô tả ở trên, vấn đề nổi trội nhất hiện nay nằm trong khả năng sẵn có các trình tự trong bộ gen cho các bộ phận cơ thể khác nhau. CSDL GenBank [2], EMBL [37] và DDBJ [38] chứa các trình tự DNA cho từng gen

mã hóa protein và sản phẩm RNA. Như nhiều CSDL trình tự protein hỗn hợp, CSDL Entrez nuleotide [39] tổng hợp các trình tự từ những CSDL thứ cấp.

Khi hoàn tất giải trình tự toán bộ bộ gen, sẽ đưa đến việc công bố những bộ gen riêng biệt tại các Site khác nhau. CSDL gen Entrez [40] gom tất cả bộ gen hoàn chỉnh của những Site vị trí riêng lẻvà hiện biểu diễn trên 1,000 bộ phận cơ thể khác nhau (Tháng 8 2000). Thêm vào đó, việc cung cấp các trình tự nucleic thô, thông tin được thể hiện ở nhiều mức độ chi tiết khác nhau bao gồm: một danh sách bộ gen hoàn chỉnh, các nhiễm sắc thể trong một bộ phận cơ thể, các quan sát chi tiết trong từng nhiễm sắc thể riêng lẻ có đánh dấu các vùng mã hóa và không mã hóa, và các gen đơn lẻ. Tại mỗi mức độ, có những thể hiện đồ họa, những phân tích tính toán, những liên kết với những phần khác nhau của Entrez. Ví dụ, những giải thích cho một gen bao gồm trình tự protein được dịch mã, phương án chỉnh thẳng cột trình tự với các gen tương tự trong các bộ gen khác và những tóm lược các đặc trưng thực nghiệm hay chức năng dự đoán. GenreCensus [41] cũng cung cấp các đề mục phân tích bệ gen trong tiến trình phát triển. CSDL này cho phép xây dựng cây phát sinh loài dựa trên điều kiện khác nhau như ribosomal RNA hay sự xuất hiện các nếp gấp trong protein. Các Site này còn cung cấp các phương án so sánh nhiều bộ gen, phân tích một bộ gen đơn lẻ và phục hồi thông tin cho từng gen riêng biệt. CSDL COG [20] phân loại protein đã mã hóa trong 21 bộ gen hoàn chỉnh trên cơ sở các trình tự tương đương. Thành viên của cùng cụm của Nhóm Orthologous, COG được mong đợi có kiến trúc 3D giống nhau và những chức năng tương đương. Khuynh hướng ứng dụng của hầu hết CSDL là dự đoán chức năng của protein không đặc trưng dựa trên tính tương đồng của chúng đối với protein đặc trưng, và cũng để nhận diện mẩu phát sinh loài của sự xuất hiện protein – ví dụ, COG được cho

được thể hiện qua hầu hết hay tất cả các bộ phận cơ thể hay chỉ trong vài loài liên quan gần.

### 4.4.Dữ liệu biểu hiện gen

Hầu hết nguồn dữ liệu mới đây đều xuất phát từ những thực nghiệm biểu hiện nhằm định lượng mức độ biểu hiện của các gen riêng lẻ.. Những thực nghiệm này đo lường số lượng mRNA hay sản phẩm protein được sản sinh bởi tế bào. Đối với vấn đề trước, có ba công nghệ chính là cDNA microarray[42-44], Affymatric GenreChip [45] và những phương pháp SAGE [46]. Phương pháp đầu tiên đo mức đô tương đối của nhiều mRNA giữa những mẩu khác nhau, trong khi hai kỹ thuật sau đó đo những mức độ tuyệt đối. Hầu hết những nỗ lực trong phân tích biểu hiện gen đều tập trung vào yeast và bộ gen người và` cho đến nay, chưa có kho lưu trữ tập trung cho dữ liệu này.

### 4.5.Tích hợp dữ liệu

Những nghiên cứu có ích nhất trong ngành sinh tin học là việc tích hợp kết quả từ nhiều nguồn dữ liệu [58]. Thí dụ, tọa độ 3D của protein thì càng hữu dụng nếu kết hợp dữ liệu về chức năng protein, sự xuất hiện trong các bộ gen khác, và sự tương tác với những phân tử khác. Theo cách này, những mẩu cá biệt của thông tin được đặt trong ngữ cảnh có đối chiếu với các dữ liệu khác. Thật không may, nó không luôn luôn dễ dàng truy cập qua những tham khảo chéo những nguồn thông tin này do sự khác biệt trong cách đặt tên và khuôn mẩu tập tin.Về cơ bản, vấn đề này thường được giải quyết bằng cách cung cấp các liên kết ngoài đến các CSDL khác, ví dụ trong PDBSUm, trang web cho cấu trúc riêng biệt trực tiếp giúp người dùng đi đến phần thích hợp trong CSDL PDB, NDB, CATH, SCOP và SWISS-PROT. Ở mức cao hơn, có những nổ lực

nhằm tích hợp truy cập chéo các nguồn dữ liệu. Một số hệ thống tiêu iểu như hệ thống truy cập trình tự SRS (Sequence Retrieval System) [59] với CSDL có cấu trúc phẳng và hỗ trợ người dùng tìm, liên kết và truy cập đến từng nucleic acid, chuỗi protein, protein motif, cấu trúc protein. Kế đến là tiện ích của CSDL Entrez [39], cho phé[ truy cập các trình tự DNA và protein, bộ gen , cấu trúc đại phân tử 3D . Khả năng tìm kiếm một gen đặc biệt trong CSDL sẽ cho phép các chuyển dịch êm ả từ bộ gen mà xuất phát, trình tự protein được mã hoá, cấu trúc của nó, thư mục tham khảo và bộ phận tương đương chcho tất cả các gen liên quan.

**4.6..Hiểu và tổ chức thông tin**

Xem xét dữ liệu, chúng ta có thể thảo luận các loại phân tích cần hướng đến. Có thể phân chia các nguồn thông tin được dùng trong những nghiên cứu theo các lĩnh vực nghiên cứu của Sinh tin học . Đối với trình tự sinh học DNA thô, các nghiên cứu tập trung vào việc phân biệt những vùng được mã hóa hoặc không mã hóa, nhận diện introns, exons và những vùng promoter cần cho việc nghiên cứu bộ gen ở dạng DNA [61,62]. Đối với trỉnh tự protein, các nổ lực phát triển thuật giải phục vụ chỉnh thẳng cột [63], tìm các vùng chức năng được bảo tồn, các đoạn lặp trong các phương án chỉnh thẳng cột, các phương pháp nghiên cứu chỉng thẳng cột trong không gian 3 chiều dùng các độ đo khoảng cách và góc, các tính toán bề mặt, hình dạng và những phân tích tương tác các protein với các đơn vị nhỏ hơn, DNA, RNA và những phân tử nhỏ hơn.

Việc tăng khả năng giải trình tự gen dẫn đến các phương pháp tính toán trên bộ gen hay bộ protein – những phân tích trên bình diện rộng của những bộ gen hoàn chỉnh và protein mà chúng mã hóa. Các nghiên cứu bao gồm đặc trưng

nội dung protein và cách thức trao đổi chất giữa bộ gen khác, quá trình nhận diện các tương tác protein, gán và dự đoán các sản phẩm gen, và những phân tích bình diện rộng của mức độ biểu hiện gen. Một vài chủ đề nghiên cứu theo hướng này sẽ được trình bày trong những phân tích ví dụ phân tích các hệ thống điều hoà phiên mã.

Những lãnh vực nghiên cứu khác cũng được đầu tư phát triển như xây dựng thư viện số chứa các tài liệu nghiên cứu về sinh tin học từ tài liệu, những phương pháp phân tích DNA trong lĩnh vực hình sự, những dự đoán cấu trúc nucleic acid, mô phòng cách thức chuyển hóa, liên kết gen cụ với những nét đặc điểm của các bệnh khác nhau.

Thêm vào đó để tìm quan hệ giữa những protein khác nhau, nhiều nhà sinh tin học đã tiến hành phân tích một loại dữ liệu để suy ra và hiểu được những quan sát cho những loại dữ liệu khác. Dùng dữ liệu trình tự hay có dữ liệu cấu trúc để dự đoán cấu trúc bậc hai hay bậc ba dựa trên các phương pháp thống kê suy diễn. Dùng dữ liệu có cấu trúc để hiểu chức năng protein. Những nghiên cứu các mối liên hệ giữa các protein tương đồng về chức năng [68,69] và những phân tích tương đương giữa những vị trí liên kết khác nhau [70]. Kết hợp với việc đo lường tính tương đương, những nghiên cứu này cho phép hiểu biết chính xác các thông tin được chuyển dịch giữa các protein tương đương.

### 4.7. Phát triển nghiên cứu ngành sinh tin học theo bề rộng và sâu

Hai nhiệm vụ chính của ngành Sinh Tin học là **tổ chức và hiểu biết** dữ liệu sinh học – sự phát triển của công nghệ sinh tin học cho phép mở rộng những phân tích sinh học theo 2 chiều, sâu và rộng.

Theo bề sâu sẽ bao gồm các nghiên cứu nhằm hiểu biết ngày càng nhiều các protein. Bắt đầu với một gen, xác định chuỗi protein, từ đó dự đoán cấu trúc của protein. Dựa vào các tính toán hình học có thể dự đoán hình dạng và bề mặt protein, mô phòng phân tử phân tử. Nhận diện liên kết, và suy đoán chức năng protein. Thực tế, những bước trung gian vẫn khó thực hiện chính xác, và cần kết hợp với những phương pháp khác để đạt kết quả mong muốn.

Theo chiều rộng sẽ bao gồm các phương pháp so sánh gen này với gen khác, protein này với peotein khác. Ban đầu là những thuật giải đơn giản được dùng để so sánh chuỗi và cấu trúc của cặp protein liên quan. Khi dữ liệu sinh học gia tăng mạnh mẽ sẽ phát sinh nhu cầu cải tiến các thuật giải có hiệu suấtr cáo để chỉng thẳng cột nhiều trình tự , trích rút mẩu chuỗi hay mẩu cấu trúc xác định họ protein, tạo cây phát sinh loài để khảo sát quá trình tiến hoá của protein. Cuối cùng, do thông tin được lưu trong CSDL lớn, công việc so sánh trở nên phức tạp hơn, đòi hỏi nhiều cải tiến trong cơ chế tổ chức và quản lý CSDL.

### 4.8. Ứng dụng kỹ thuật tin học

Nhiều lĩnh vực Sinh Tin học đòi hỏi các kỹ thuật tin học khác nhau: đối với tổ chức dữ liệu, CSDL sinh học những sử dụng các tập tin phẳng đơn giản. Tuy nhiên khi gia tăng số lượng thông tin, các CSDL quan hệ với giao diện Web sẽ ngày càng phổ biến, những kỹ thuật mới cần phát triển bao gồm phương pháp so sánh chuỗi, thuật giải chỉnh thẳng cột, nhận diện đoạn lặp ( motif) và các phương pháp máy học, phân cụm và kỹ thuật khai thác dữ liệu. Việc phân tích có cấu trúc 3D bao gồm tính toán hình học Euclid kết hợp với ứng dụng cơ bản của hóa lý, thể hiện đồ họa của bề mặt và hình khối, và sự so sánh cấu trúc và phương pháp hợp 3D. Trong mô phòng phân tử, cơ chế Newton, cơ chế định

lượng, cơ chế phân tử, những tính toán tĩnh điện đã được áp dụng. Trong nhiều lãnh vực này, phương pháp tính toán phải được kết hợp những phân tích thống kê tốt để cung cấp sư các số liệu và kết quả có ý nghĩa tốt.

## 6. Điều hoà phiên mã – một nghiên cứu trong sinh học

Những protein gắn kết DNA đóng vai trò trọng tâm trong tất cả các mặt của hoạt động di truyền trong bộ phân cơ thể, chúng tham gia vào những quá trình như là sự phiên mã, đóna'gói, tái sắp xếp, sao lại hay sửa chữa. Phần nay, trình bàu các nghiên cứu nhằm tìm hiểu sự điều hoà phiên mã trong những bộ phận cơ thể khác nhau. Qua ví dụ này, chúng ta sẽ chứng minh được các kỹ thuật sinh tin học đã được dùng để nâng cao kiến thức về hệ thống sinh học và cũng minh họa những ứng dụng thực tế của những lãnh vực khác nhau đã được phác thảo trước đây.

Trước tiên cần xem xét đến những phân tích cấu trúc về cách protein gắn kết DNA nhận diện trình tự cơ bản. Sau đó sẽ lược qua những nghiên cứu bộ gen đặc trưng những yếu tố phiên mã trong các bộ phận cơ thể khác nhau, và những phương pháp được dùng để nhận diện vị trí kết hợp điều hoà trong vùng ngược dòng. Cuối cùng là một khái quát những phân tích biểu hiện gen đã được hướng dẫn gần đây vả đề nghị sử dụng trong tương lai Tất cả những kết qủa mô tả được tìm thấy nhờ vào các nghiên cứu kỹ thuật tính toán.

### 6.1. Những nghiên cứu có cấu trúc

Vào tháng 8 2000, có 379 cấu trúc của phức hợp protein DNA trong PDB. Những phân tích của cấu trúc cung cấp các hiểu biết giá trị theo các nguyên lý trong

hóa học lập thể của việc kết hợp bao gồm cách nhận dạng các trình tự cơ dở, cách sửa đổi các liên kết của  cấu trúc DNA .

Nguyên tắc phân loại có cấu trúc của protein gắn kết DNA, tương tư đều thể hiện trong SCOP và CATH, đầu tiên được đề xuất bởi Harrison [72] và cập nhật một cách định kỳ để chứa những cấu trúc mới khi chúng được giải quyết [73]. Sự phân loại gồm có hệ thống hai lớp: mức đầu tiên gom protein thành 8 nhóm mà chia sẻ tổng đặc trưng có cấu trúc cho DNA gắn kết, và thứ hai gồm 54 họ protein mà tương đồng với nhau một cách có cấu trúc. Sự lắp ráp của một hệ đơn giản hóa sự so sánh của những phương pháp kết hợp khác nhau; chúng làm nổi bật tính đa dạng của protein –những cấu trúc hình học DNA phức được tìm thấy trong tự nhiên, nhưng cũng nhấn mạnh điểm quan trọng của những giao thức giữa hình xoắn ∝ và đường nét DNA chính, phương thức chính của việc kết hợp hơn phân nữa họ protein. Trong khi số những cấu trúc thể hiện trong PDB không phản ánh một cách thiết yếu điều quan trọng liên quan của những protein khác nhau trong tế bào, rõ ràng là hình xoắn-sang-hình xoắn, zinc-coordinating và leucine zipper motif được sử dụn lập lại. Điều này cung cấp khuôn khổ rắn chắc có mặt hình xoắn ∝ trên bề mặt của những protein dẫn xuất một cách có cấu trúc. Ở mức tổng quát, có thể làm nổi bật những điểm khác nhau giữa những phạm vi yếu tố sao chép mà “chỉ” gắn kết DNA từ mặt đơn và vị trí thành đường rãnh để giao tiếp với cạnh cơ bản. Sau đó nói chung phủ lớp nền, dùng mạng phức tạp của cấu trúc thế hệ thứ hai và lặp lại.

Tập trung trên protein với hình xoắn ∝, cấu trúc biểu diễn nhiều sự biến đổi, cả chuỗi acid amino và hình học chi tiết. Chúng tiến triển một cách độc lập theo những đòi hỏi trong ngữ cảnh mà chúng được tìm thấy. Khi đạt được điều chỉnh chặt chẽ giữa hình xoắn ∝ và đường nét chính, có đủ tính linh hoạt cho phép cả

protein và DNA để chấp nhận hình thể cấu tạo phân biệt. Tuy nhiên, một vài nghiên cứu đã phân tích những hình học kết hợp của hình xoắn ∝ đã trình bày rằng hầu hết chấp nhận hình thể hoàn toàn giống nhau bất chấp họ protein. Nói chung, chúng được thêm vào đường nét chính, với chiều dài trục xem như song song với độ dốc phác họa bởi DNA trụ cột.

Hầu hết bắt đầu N điểm dừng trong đường nét chạy dài, hoàn tất 2 đến 3 lần theo khoảng cách tiếp xúc của acid nucleic. [75,76].

Cho sự định hướng kết hợp tương đương, bất ngờ phát hiện những giao tiếp giữa mỗi vị trí acid amino dọc theo hình xoắn ∝ và nucleotides trên DNA thay đổi có thể xem như giữa họ protein khác nhau. Tuy nhiên, bằng việc phân loại acid amino ứng những độ dài của chuỗi cạnh, theo hợp thức hóa những mẫu giao tiếp khác nhau. Luật của những giao tiếp được dựa trên giả thuyết đơn giản cho một vị trí thặng dư được cho trên hình xoắn ∝ trong những hình thể tương tự, acid amino nhỏ giao tiếp với nucleotides gần theo khoảng cách và acid amino lớn với những giao tiếp đó thì xa hơn [76,77]. Những nghiên cứu tương đương cho việc kết hợp bởi những motif có cấu trúc khác giống β-hairpins, cũng được quản lý[78]. Khi xem xét những giao tiếp này, điều quan trọng là vùng khác nhau của bề mặt protein cũng cung cấp những bề mặt chung với DNA.

Điều này mang đến cái nhìn mức nguyên tử những giao tiếp giữa các cặp acid amino riêng biệt. Những phân tích như vậy dựa trên giả thiết mà sự cân đối có ý nghĩa của kết hợp DNA được hợp thức hóa bởi mã phổ quát của sự nhận diện giữa acid amino và nền cơ bản, i.e phần còn lại protein chắc chắn tốt nhất giao tiếp với nucleotide riêng biệt bất chấp loại protein-DNA phức [79]. Những nghiên cứu được xem xét sự liên kết hydrô, những tiếp xúc Van der Waals, sự liên kết nước làm trung gian [80-82]. Những kết qủa chỉ ra rằng khoảng 2/3 của

những tác động qua lại thì cùng với DNA trụ cột và vai trò chính của chúng là một trong sự ổn định độc lập chuỗi. Trái lại, những tác động qua lại với những cơ bản hiển thị một vài quyền ưu tiên mạnh, bao gồm những tác động qua lại của arginine hay lysine với guanine, asparagine hay glutamine với adenine và threonine với thymine. Những quyền ưu tiên như vậy được giải thích qua sự kiểm tra hóa học lập thể của chuỗi cạnh acid amino hay những cạnh cơ bản. Làm nổi bật thì cũng phức tạp hơn những kiểu tác động qua lại nơi acid amino đơn tiếp xúc đồng thời với nhiều hơn một bước đơn, theo cách đó nhận diện DNA ngắn. Những kết quả này đề nghị rằng những nét đặc trưng phổ quát, một được quan sát tất cả protein- DNA phức, quả thực tồn tại. Tuy nhiên, nhiều tác động qua lại bình thường được xem như không cụ thể như là những tác động đó với DNA trụ cột, có thể cũng cung cấp đặc trưng phụ thuộc ngữ cảnh mà chúng được tạo ra.

Trang bị một sự hiểu biết cấu trúc protein, những motif gắn kết DNA và chuỗi cạnh hóa học lập thể chuỗi cạnh, một ứng dụng chính là sự dự đoán của việc gắn kết bởi những protein đã biết để chứa những motif riêng biệt hoặc những protein đó với cấu trúc đã được tháo gỡ trong hình dạng không phức tạp. Nói chung hầu hết là những dự đoán chonhững tác động qua lại đường nét chính của hình xoắn α - chuỗi  acid amino đã cho mà chuỗi DNA nhận diện [77,83]. Trong một cách tiếp cận khác, kỹ thuật giả lập phân tử đã được dùng cắt ngắn toàn bộ protein và DNA trên nguyên tắc cơ bản của những tính toán quanh 2 phân tử [84,85]

Hai phương pháp chỉ đi đến những thành công giới hạn bởi vì những trường hợp đơn giản như kết hợp hình xoắn α, có nhiều những yếu tố khác phải được xem xét. So sánh giữa cấu trúc acid nucleic kết hợp hay không kết hợp chỉ ra rằng việc gắn kết DNA là một đặc trưng chung của những liên hợp được hình thành

bởi những yếu tố sao chép [74,86]. Điều này và những yếu tố khác như tĩnh điện và sự tác động lấy nước làm trung gian giúp việc nhận diên gián tiếp của chuỗi nucleotide, mặc dù chúng chưa hiểu rõ. Như vậy, những luật chi tiết cho sư gắn kết DNA sẽ là họ cụ thể nhưng dưới xu hướng những tác động qua lại arginine-guanine.

**Những nghiên cứu hệ gen.**

Bởi tính phong phú của dữ liệu hóa sinh luôn sẵn có, những nghiên cứu gen trong sinh tin học tập trung trên mô hình cơ thể và sự phân tích hệ thống điều tiết không là ngoại lệ. Sự đồng nhất những yếu tố sao chép trong hệ gen có tính không thay đổi phụ thuộc vào chiến lược nghiên cứu sự tương đồng thừa nhận mối liên hệ chức năng và tiến triển giữa những protein tương đương. Trong E.coli, những nghiên cứu ước lượng khoảng từ 300 đến 500 những điều tiết sao [87] và PEDANT [88], một CSDL của những chức năng gen kết nối tự động, chỉ ra rằng điển hình 2-3% gen prokaryotic và 6-7% gen eukaryotic  bao gồm những protein gắn kết DNA. Khi sự phân công chỉ hoàn tất 40-60% hệ gen như tháng 8 năm 2000, những minh họa này hầu như trên dữ liệu thực. Tuy nhiên, chúng đại diện cho số lượng lớn protein và có nhiều điều chỉnh sao chép trong eukaryotes hơn loại khác. Điều này không có gì ngạc nhiên, xem xét những cơ quan đãphát triển một cơ chế sao chép phức tạp có liên quan.

Từ kết luận của những nghiên cứu có cấu trúc, chiến lược tốt nhất cho việc đặc trưng hóa gắn kết DNA của những yếu tố sao chép giả định trong mỗi hệ gen là để nhóm chúng lại bởi tính tương đồng và phân tích những họ riêng biệt. Sự phân loại như vậy được cung cấp trong CSDL chuỗi thế hệ thứ hai mô tả trước và sự phân loại đó đặc biệt hóa trong những điều tiết protein như là RegulonDB

[89] vàTRANSFAC [90]. Cùng sử dùng lớn hơn là sự cung cấp của sự phân công có cấu trúc đối với protein; cho một một yếu tố sao chép, nó hữu dụng để biết motif có cấu trúc dùng cho việc gắn kết, như vậy cung cấp sự hiểu biết tốt hơn làm thế nào nó nhận biết chuỗi gốc. Hệ gen có cấu trúc qua sinh tin học ấn định những cấu trúc đối với sản phẩm protein của hệ gen bởi việc chứng minh tính tương đồng đối những protein của cấu trúc đã biết [91]. Những nghiên cứu này chỉ ra rằng những yếu tố sao chép prokaryotic hầu như thường chứa motif hình xoắn-sang-hình xoắn [87,92] và yếu tố eukaryotic chứa loại homeodomain hình xoắn-sang-hình xoắn, zinc finger hay leucine zipper motif. Từ những phân loại protein của mỗi gen, rõ ràng rằng những kiểu khác nhau của protein điều tiết khác biệt nhiều và những họ khác biệt kích thước 1 cách đáng kể. Một nghiên cứu bởi Huynen và van Nimwegen [93] chỉ ra rằng những thành viên của một họ đơn có những chức năng giống nhau, nhưng đòi hỏi những chức năng này biến đổi vượt thời gian, vì vậy làm những thể hiện của mỗi hệ gen trong hệ gen.

Hầu như gần đây, dùng sự kết hợp của chuỗi và dữ liệu có cấu trúc, kiểm tra sự giao hợp của chuỗi acid amino giữa protein gắn kết DNA liên quan và ảnh hưởng  sự biến đổi có trên việc nhận diện chuỗi DNA. Những họ có cấu trúc mô tả ở trên đã được mở rộng bao gồm protein có liên hệ bởi sư tương đồng chuỗi, nhưng toàn bộ cấu trúc không được tìm ra. Ngoài ra, những thành viên của cùng một họ thì tương đồng, và dẫn xuất từ tổ tiên chung.

Những giao kết acid amino được tính toán cho những liên kết đa chuỗi của mỗi ho [94]. Nói chung, vị trí liên kết tác động với DNA được bảo quản tốt hơn phần còn lại của bề mặt protein, mặc dù những mẩu chi tiết của việc giao kết là hoàn toàn phức tạp. Những phần còn lại tiếp xúc với DNA trụ cột được bảo quản cao

trong tất cả họ protein, cung cấp một sự ổn định những tác động chung đối với tất cả protein tương đồng. Những giao cấu của những vị trí liên kết mà tiếp xúc những cơ sở, và nhận diện chuỗi DNA, thì càng phức tạp và hợp lý hóa bởi định nghĩa mô hình 3 lớp cho DNA gắn kết. Đầu tiên, họ protein gắn kết một cách không cụ thể thường chứa một vài phần còn lại tiếp xúc cơ bản được bảo quản; không ngoại lệ, những tác động qua lại được là bởi đường nét nhỏ mảở đó có sự phân biệt nhỏ giữa những loại cơ bản. Những tiếp xúc nói chung để ổn định sự biến dạng trong cấu trúc acid nucleic, đặc biệt trong việc nới rộng đường nét nhỏ DNA. Lớp thứ hai bao gồm những họ thành viên nhằm làm cho tất cả chuỗi nucleic đích giống nhau; Ở đây, những vị trí tiếp xúc cơ bản được bảo quản tuyệt đối hay cao hơn cho phép những protein liên quan đạt chuỗi giống nhau. Lớp thứ ba, đáng chú ý, bao gồm những họ mà trong việc gắn kết cụ thể những thành viên khác nhau gắn kết những chuỗi cơ bản phân biệt. Phần còn lại của protein này thường xuyên trải qua đột biến, và những thành viên họ hàng có thể dẫn xuất từ những họ con tương ứng chuỗi acid amino tại vị trí tiếp xúc cơ bản; những thành viên đó trong cùng một họ con được dự đoán bởi gắn kết cùng chuỗi DNA và những thành viên của những họ con khác nhau gắn kết chuỗi phân biệt.  Tổng quát, những họ con tương ứng tốt với những chức năng của protein và những thành viên của cùng một họ con được tìm để điều tiết cách thức sao chép tương đương. Những phân tích phối hợp của chuỗi và dữ liệu có cấu trúc mô tả bởi nghiên cứu này cung cấp một sự hiểu biết sâu sắc làm thế nào nền tảng gắn kết DNA tương đồng đạt được những đặc trưng khác nhau bởi biến đổi những chuỗi acid amino của chúng. Trong việc làm như vậy, protein tiến triển những chức năng phân biệt, cho phép một cách có cấu trúc những yếu tố sao chép điều chỉnh biểu thức của gen khác nhau. Như vậy, tính phong phú tương đối của những họ điều tiết sao chép trong hệ gen phụ thuộc vào tầm quan

trọng của một chức năng protein đặc trưng, nhưng cũng có khả năng thích ứng của những motif gắn kết DNA để nhận diện chuỗi nucleic phân biệt. Trong trường hợp này có vẻ như được cung cấp bởi những motif gắn kết đơn giản, như zincd fingers.

Những kiến thức của những điều tiết sao chép được chứa trong mỗi cơ quan, và một sự hiểu biết làm thế nào nhận diên chuỗi DNA, thật thú vị để tìm những vị trí gắn kết then chốt của chúng với chuỗi gen, hầu hết những phân tích liên qua dữ liệu kết hợp trên những vị trí gắn kết được biết qua thực nghiệm cho những protein đặc trưng và tạo nên chuỗi liên ứng kết hợp chặt chẽ bất kỳ sự biến đổi trong nucleic. Những vị trí thêm vào được tìm thấy bởi việc quản lý so từ tìm kiếm trên toàn hệ gen và đánh dấu những vị trí ứng viên bởi sự tương đương [96-99]. Không ngạc nhiên, hầu hết của những vị trí được dự đoán được tìm trong những vùng mã hóa của DNA [96] và kết quả của những nghiên cứu thường được hiện diện trong CSDL như RegulonDB [89]. Hướng tiếp cận tìm tiếp liên ứng thường được thực thi bởi những nghiên cứu gen tương đối tìm kiếm ngược dòng những vùng của những gen orthologous trong những cơ quan liên quan một cách chặt chẽ. Qua hướng tiếp cận như vậy, nó phát hiện ra rằng có ít nhất 27% của motif điều tiết DNA E.coli đã biết được bảo quản trong một hay nhiều vi khuẩn liên quan cách xa nhau [100].

Sự phát hiện những vị trí điều tiết trong eukaryotes đặt ra một vấn đề khó hơn bởi vì khuynh hướng chuỗi liên ứng cho là ngắn nhiều hơn, biến đổi, và phân tán trên khoảng cách rất rộng. Tuy nhiên, những nghiên cứu khởi đầu trong S. cerevisiae đã cung cấp một quan sát đáng chú ý cho protein GATA trong sự điều tiết trao đổi chất nitrogen. Trong khi chuỗi liên ứng 5 cặp cơ bản GATA được tìm thấy hầu hết mỗi nơi trong hệ gen, một vị trí gắn kết đơn lập thiếu sử dụng

chức năng điều tiết [101]. Những nét đặc trưng như vậy của hoạt tính GATA xuất phát từ sự lặp lại của chuỗi liên ứng với những vùng đối nghịch của những gen được kiểm soát trong nhiều sao chép. Một nghiên cứu khởi đầu sử dụng quan sát này để dự đoán những vị trí điều tiết mới, bởi việc tìm kiếm trên oligonucleotide, được biểu trưng trong những vùng mã hóa của men và hệ gen ký sinh [102,103].

Việc phát hiện những vị trí điều tiết gắn kết, có vấn đề của việc định nghĩa gen thực sự đã được điều tiết, thuật ngữ chung là regulon. Nói chung, vị trí gắn kết giả sử rằng được định vị trực tiếp đối nghịch của regulons; tuy nhiên, có những vấn đề khác nhau kết hợp với thừa nhận này phụ thuộc vào cơ quan. Đối với prokaryotes, nó làm rắc rối thêm bởi sự hiện diện của operons; định vị gen đã điều tiết với một operon khi nó có thể nằm trong một vài gen xuôi dòng của chuỗi điều tiết là rất khó. Khó có thể dự đoán tổ chức của operons [104], đặc biệt để định nghĩa gen mà được tìm thấy tại đầu, và thường có một sự thiếu hụt của sự bảo quản lâu dài trong trình tự gen giữa những tổ chức liên quan [105]. Vấn đề trong eukaryotes thậm chí rất gay gắt; những vị trí điều tiết thường hành động theo hai hướng, những vị trí gắn kết thường cách xa regulons, và việc điều tiết sao chép là kết qủa của hành động kết hợp bởi nhiều yếu tố sao chép trong một kiểu tổ hợp.

Mặc dù, những vấn đề thành công trong việc thừa nhận cách thức điều tiết sao chép của những hệ thống đặc trưng hóa tốt như hệ thống phản ứng va chạm hơi nóng [99].Thêm vào đó, để kiểm tra một cách thực nghiệm bất kỳ những dự đoán, hầu hết dùng dữ liệu diễn giải gen.

**Những nghiên cứu diễn giải gen**

Nhiều nghiên cứu tập trung trên phát minh những phương pháp để phân cụm gen bởi sự tương đồng trong mô tả sơ lược diễn giải. Điều này để định rõ protein được biểu diễn cùng nhau dưới điều kiện tế bào khác nhau. Ngắn gọn, những phương pháp chung nhất đang phân cụm có thứ bậc. Những phương pháp có thứ bậc nguồn gốc dẫn xuất từ những thuật toán để xây dựng cây phát sinh loài, nhóm gen trong kiểu từ dưới lên; gen với những mô tả sơ lược diễn giải tương tự nhất được phân cụm đầu tiên, và những gen nàyvới nhiều mô tả sơ lược biến đổi được bao gồm lặp đi lặp lại [106-108]. Trái lại, bảng đồ tự tổ chức [109-110] và phương pháp K-means [111] dùng hướng tiếp cận trên xuống để dùng định nghĩa trước số phân cụm cho tập dữ liệu. Những phân cụm khởi đầu gán trị ngẩu nhiên, và những gen được nhóm lặp đến khi chúng được phân cụm một cách tối ưu.

Những phương pháp đã cho liên quan đến dữ liệu diễn giải thuộc tính khác như cấu trúc, chức năng và sự định vị tế bào con của mỗi sản phẩm gen. Ánh xạ những thuộc tính này cung cấp một sự hiểu biết bên trong những đặc trưng của protein mà được diễn giải cùng với nhau, và cũng đề nghị những kết luận đáng quan tâm về toàn thể hóa sinh của tế bào. Trong men, những protein ngắn hơn có khuynh hướng được diễn giải ở mức cao hơn những protein dài hơn, có lẽ bởi vì sự không ràng buộc liên hệ mà chúng được tạo ra [112]. Nhìn vào nội dung của acid amino, những gen được diễn giải ở mức cao nói chung giàu alanine và glycine, và làm suy yếu trong asparagine; những điều này được suy xét để phản ánh những đòi hỏi của acid amino dùng trong cơ thể; nơi sự tổng hợp của alinine và glycine thì cao hơn asparagine. Xoay cấu trúc protein, mức diễn giải của TIM barrel và những nhánh hydrolase NTP là cao nhất, trong khi mức diễn giải đó

cho những nhánh leucine zipper, zinc finger và màng chuyển dịch chứa helix là thấp nhất. Điều này liên quan đến những chức năng kết hợp với những nhánh này; những vấn đề trước liên quan đến cách thức trao đổi chất và vấn đề sau liên quan  những tiến trình chuyển tín hiệu hay chuyền tải [113]. Điều này phản ánh trong mối liên hệ với sự định vị tế bào con của protein, nơi diễn giải của những protein cytoplasmic là cao, nhưng những protein nhân và màng có khuynh hướng trở nên thấp.

Những mối quan hệ càng phức tạp cũng được đánh giá. Những hiểu biết quy ước các sản phẩm gen tác động lẫn nhau có thể có những sơ lược diễn giải tương tư hơn nếu chúng không tác động [116,117]. Tuy nhiên, một nghiên cứu gần đây chỉ ra rằng mối liên hệ này không quá đơn giản [118]. Trong khi những sơ lược diễn giải thì tương đương những sản phẩm gen kết hợp thường xuyên, ví dụ trong một đơn vị con ribosomal, những mô tả khác biệt chủ yếu đối với những sản phẩm chỉ kết hợp ngắn, bao gồm những thuộc loại đó đối với cách thức trao đổi chất như nhau.

Theo những mô tả ở trên, một trong những tác động dẫn xuất chính dưới sự phân tích diễn giải là để phân tích những dòng tế bào ung thư [119]. Nói chung, những dòng tế bào khác nhau (ví dụ tế bào epithelial và ovarian) có thể được phân biệt trên cơ bản của những mô tả diễn giải, và những mô tả sơ lược này được duy trì khi tế bào được chuyển dịch từ một môi trường *in vivo* đến một môi trường *in vitro* [120]. Cơ bản cho những khác biệt vật lý rõ ràng trong biểu thức gen cụ thể; ví dụ mức độ diễn giải sản phẩm gen cần thiết cho tiến trình qua chu trình tế bào. Đặc biệt gen ribosomal tương quan với nhau tốt trong những biến đổi trong tỷ lệ phát triển tế bào. Sự phân tích so sánh có thể được mở rộng đến tế bào u bướu mà nguyên nhân cơ bản của ung thư có thể được hé mở bởi vùng rất nhỏ

của sự biến đổi sinh học so sánh với những tế bào thường. Ví dụ, trong tế bào vú, gen liên quan đến sự phát triển tế bào và cách thức chuyển đổi luồng dấu hiệu IFN điều tiết được tìm thấy để được làm điều tiết [52,121]. Một trong những khó khăn trong trị bệnh ung thư mục tiêu chữa bệnh 1 cách cụ thể đối với loại ung bướu phát sinh bệnh khác biệt, để tối ưu hóa tính hiệu quả và cực tiểu hóa đặc tính độc. Như vậy, những cải thiện trong phân loại ung thư, tập trung để thúc đẩy trong trị bệnh ung thư. Mặc dù những khác biệt giữa hình thức của bệnh ung thư – ví dụ những lớp con của bệnh bạch cầu cấp tính được chính thức hóa, và chưa thể thiết lập một chẩn đoán buồng trứng trên nền tảng một kiểm tra đơn giản. Trong một nghiên cứu gần đây bệnh bạch cầu myeloid cấp tính được phân biệt thành công dựa trên những mô tả diễn giải của những tế bào này [53]. Khi những phương pháp không đòi hỏi kiến thức sinh học của những căn bệnh, nó cung cấp một chiến lược chung cho việc phân loại tất cả các loại bệnh ung thư.

Rõ ràng, một khía cạnh then chốt của việc hiểu biết dữ liệu diễn giải nằm trong sự hiểu biết nền tảng của sự điều tiết chuyển dịch. Tuy nhiên, sự phân tích trong vùng này vẫn bị giới hạn đến những phân tích sơ bộ của mức diễn giải trong men mutants thiếu thành phần then chốt của sự sao chép khởi đầu phức tạp.

**"... nhiều những ứng dụng thực nghiệm..."**

Ở đây, chúng ta mô tả một vài những điểm chính dùng sinh tin học.

**Tìm những tương đồng**

Như mô tả ở trước, một trong những ảnh hưởng dẫn xuất dưới sinh tin học là việc tìm kiếm tính tương đồng giữa phân tử sinh học. Ngoài việc cho phép cơ quan có hệ thống của dữ liệu, sự đồng nhất những tương ứng protein có một số sử dụng thực nghiệm trực tiếp. Chuyển dịch những thông tin giữa protein liên quan là rõ

ràng. Ví dụ, cho một protein kém đặc trưng, nó có thể tìm sự tương đồng được tìm hiểu tốt với sự thận trọng, ứng dụng vào kiến thức trước và sau. Đặc biệt với dữ liệu có cấu trúc, mô hình lý thuyết của protein được dựa trên kinh nghiệm những cấu trúc giải quyết của những sự tương đồng gần [123]. Những kỹ thuật tương tự dùng trong sự nhận diện nhánh trong những dự đoán cấu trúc thế hệ thứ hai phụ thuộc vào việc tìm kiếm những cấu trúc tương đồng biệt lập và kiểm tra dự đoán có thể làm một cách mạnh mẽ hay không [124]. Nơi mà hóa sinh hay dữ liệu có cấu trúc thiếu, những nghiên cứu có thể làm ở những cơ quan có mức thấp giống như men và những kết quả ứng dụng cho những sự tương đồng trong những cơ quan có mức cao như con người mà những thử nghiệm càng đòi hỏi hơn.

Một hướng tiếp cận tương đương cũng được dùng trong hệ gen. Tìm sự tương đồng được dùng rộng để thừa nhận những vùng mã hóa trong chuỗi gen mới và dữ liệu chức năng thường được chuyển dịch để chú giải gen rieng biệt. Trên diện rộng, đơn giản hóa vấn đề hiểu biết những gen phức tạp bằng cách phân tích tổ chức đơn giản đầu tiên và sau đó ứng dụng những yếu tố cơ bản giống nhau đến phức tạp hơn – điều này là một nguyên nhân tại sao những dự án hệ gen có cấu trúc sớm tập trung vào Mycoplasma gentitalium [91].

Trớ trêu, ý tưởng giống nhau có thể ứng dụng ngược lại. Tác dụng thuốc protein nhanh chóng được khám phá bởi việc kiểm tra những tương đồng của những protein vi khẩn thiết yếu đang nhầm lẫn trong loài người hay không. Trong phạm vi nhỏ hơn, những khác biệt cấu trúc giữa những protein tương đương có lẽ được khai thác để tạo phân tử thuốc mà đặc biệt gắn kết với một cấu trúc nhưng không với cấu trúc khác.

**Định lượng dược phẩm có chừng mực**

Một trong những ứng dụng y khoa sớm nhất của sinh tin học có sự trợ giúp định lượng dược phẩm chừng mực. Hình 2, định lượng hướng tiếp cận nêu chung, lấy gen MLH1 như một ví dụ tác dụng dược phẩm. MLH1 là gen người mã hóa protein sữa chữa một ghép đôi không khớp (mmr) được định vị trên nhánh ngắn của nhiễm sắc thể 3[125]. Qua sự phân tích nối kết và tính tương đồng của nó thành gen mmr trong chuột, gen liên quan trong nonpolyposis colorecta cancer 126]. Cho chuỗi nucleotide, chuỗi acid amino có khả năng của protein được mã hóa có thể được xác định dùng phần mềm sao chép. Kỹ thuật tìm kiếm chuỗi có thể sau đó được dùng để tìm những tương đồng trong cơ quan kiểu mẫu, và dựa trên sự tương đương chuỗi. Nó có thể kiểu mẫu cấu trúc của protein người trên những cấu trúc đặc trưng thực nghiệm. Cuối cùng, thuật toán cắt ngắn có thể thiết kế phân tử gắn kết cấu trúc kiểu mẩu, dẫn đến cách thức cho những thử nghiệm hóa sinh để kiểm tra hoạt động sinh học trên protein thực sự.

**Những khảo sát trên diện rộng**

Dù CSDL có thể lưu trữ hiệu quả tất cả thông tin liên quan đến gen, những cấu trúc và tập diễn giải, nó hữu dụng để làm ngưng tụ tất cả thông tin này theo chiều hướng và mặt có hiểu biết mà người dùng có thể hiểu dễ dàng. Những khái quát rộng có thể giúp nhận diện lãnh vực quan tâm cho những phân tích chi tiết hơn, và thay thế những quan sát mới trong ngữ cảnh riêng. Điều này cho phép ta hiểu chúng đặc biệt trong bất cứ cách thức nào hay không.

Qua những khảo sát diện rộng, điều có thể đề cập là số sự tiến triển, những câu hỏi hóa sinh và lý sinh. Ví dụ, là những nhánh protein cụ thể kết hợp với những nhóm phát sinh loài? Sự phổ biến thì khác biệt những nhánh với những cơ quan

đặc thù như thế nào? Và những nhánh chia sẻ giữa những cơ quan liên quan ở mức độ gì? Điều này mở rộng việc chia sẻ những đo lượng song song của sự liên quan họ hàng dẫn xuất từ cây phát sinh loài truyền thống? Những nghiên cứu khởi đầu chỉ ra rằng tính thường xuyên của những nhánh khác biệt lớn giữa những tổ chức và sự chia sẻ nhánh giữa những tổ chức làm nên thực sự theo sự phân loại phát sinh loài truyền thống. Chúng ta có thể tác động lên dữ liệu trên chức năng protein; cho những nhánh protein đặc trưng mà thường liên quan đến những chức năng hóa sinh đặc trưng [68,69], những việc tìm thấy này nổi bật tính đa dạng của cách thức trao đổi chất trong những tổ chức khác nhau [20,105].

**Hình 2**

Như đã bàn ở trên, một trong những tài nguyên mới hấp dẫn nhất của thông tin gen là dữ liệu diễn giải. Kết hợp thông tin diễn giải với sự phân loại có cấu trúc và chức năng của protien có thể đòi hỏi sự xuất hiện cao của một nhánh protein trong hệ gen thì biểu thị mức diễn giải cao hay không [112]. Xem xét dữ liệu trong những khảo sát diện rộng gồm những hạn định tế bào con của những protein và những tác động của chúng với nhau [127-129]. Trong những nối kết dữ liệu có cấu trúc, bắt đầu biên dịch một bản đồ tất cả protein tác động qua lại trong một tổ chức.

**Những ứng dụng xa hơn trong khoa học y học**

Những ứng dụng gần đây trong khoa học y học tập trung trên việc phân tích diễn giải gen [130]. Điều này luôn liên quan đến việc biên dịch dữ liệu diễn giải tế bào tác động bởi những căn bệnh khác nhau [113], ví dụ bệnh ung thư [53,132,133] và ateruisckerosis [134], và việc so sánh kiểm tra ngăn chặn mức diễn giải bình thường. Sự đồng nhất gen mà được diễn giải khác nhau trong

những tế bào bị tác động cung cấp cơ sở cho việc giải thích những nguyên nhân của căn bệnh và làm nổi bật tác dụng thuốc tiềm ẩn. Dùng tiến trình mô tả trong hình 2, tiến trình có thể phác thảo những hợp chất mà gắn kết protein diễn giải, hay có lẽ quan trọng hơn, việc điều tiết sao chép gây ra những thay đổi ơ mức diễn giả. Cho một hợp chất đầu, những thực nghiệm mảng nhỏ có thể được dùng để định giá những phản ứng đối với sự can thiệp dược lý, [135,136] và cũng cung cấp những bài kiểm tra sớm để dò tìm hay dự đoán toxicity của dược phẩm thử nghiệm.

Những cải tiến xa hơn nữa trong sinh tin học kết hợp với hệ gen thực nghiệm cho cá nhân được dự đoán làm nên thay đổi lớn trong tương lai của việc chăm sóc sức khỏe. Một sự kiện bắt đầu với trạm di truyền để truy cập tính nhạy cảm hay sự miễn dịch từ những căn bệnh cụ thể và gen sinh bệnh. Với những thông tin này, sự kết hợp duy nhất của vaccine có thể kê đơn, sự thiếu trung tâm sức khỏe của việc điều trị không giá trị và việc đoán trước phản ứng dữ dội của những căn bệnh sau đó trong cuộc sống. Kế tiếp thời gian sống thường ngày có thể dẫn đến hướng dẫn cho lượng chất dinh dưỡng và sự phát hiện căn bệnh [137].

Thêm vào đó những điều trị dựa vào dược phẩm có thể biến đổi đặc biệt đến bệnh nhận và căn bệnh, như vậy việc cung cấp quá trình diễn biến có hiệu lực nhất của dược phẩm với những tác dụng phụ tối thiểu [138]. Với tốc độ phát triển hiện nay việc chăm sóc sức khỏe hiện hữu sẽ ất thuận tiện hơn trong tương lai.

**Kết luận**

Với dữ liệu tràn ngập hiện nay, phương pháp tính toán trở nên rất cần thiết đối với nghiên cứu sinh học. Khởi đầu được phát triển cho những phân tích của chuỗi

sinh học, sinh tin học ngày nay trên phạm vi rộng bao gồm sinh học cấu trúc, hình học và những nghiên cứu diễn giải gen. Trong phần tổng quan này, giới thiệu chung và khái quát hiện trạng lãnh vực. Đặc biệt, bàn bạc loại thông tin sinh học và CSDL dùng chung kiểm tra một vài nghiên cứu được quản lý – với tham khảo đến hệ thống điều tiết sao chép – cuối cùng xem qua một vài ứng dụng thực nghiệm của lành vực này.

Hai hướng tiếp cận chính làm cơ sở tất cả nghiên cứu trong sinh tin học. Đầu tiên là việc so sánh và nhóm dữ liệu tương ứng với sự tương tư mang ý nghĩa sinh học và thứ hai mà sự phân tích một loại dữ liệu để suy ra và hiểu sự quan sát cho những dữ liệu khác. Những hướng tiếp cận được phản ánh những mục tiêu chính của lãnh vực này, mà được hiểu và tổ chức hóa thông tin kết hợp phân tử sinh học trên diện rộng. Như một kết qủa, sinh tin học không chỉ cung cấp chiều sâu rộng hơn đối với nghiên cứu sinh học, mà còn thêm chiều rộng. Điều này có thể kiểm tra những hệ thống cá nhân 1 cách chi tiết và cũng so sánh chúng với những điều liên quan theo trình tự những vấn đề chung được ứng dụng qua nhiều hệ thống và làm nổi bật những đặc trưng không thường xuyên duy nhất đối một số khác.