

Lời tựa

Khoa học ngày càng mang tính liên ngành, và sinh học đang trở thành một nhánh lớn của khoa học dữ liệu hiện đại. Chỉ trong khoảng hai mươi năm trở lại đây, kỹ thuật sinh học đã phát triển với tốc độ chưa từng thấy, với nhiều công nghệ thông lượng cao (*high-throughput technologies*) cho phép đọc thông tin phân tử ở quy mô lớn. Ví dụ điển hình là chi phí giải trình tự một hệ gen người đã giảm từ hàng trăm triệu đô la xuống còn vài trăm đô la. Cùng với lượng dữ liệu đó, ranh giới giữa các ngành học cũng ngày càng mờ đi. Một luận văn thạc sĩ ngành sinh học ngày nay có thể chứa nhiều dòng lệnh hơn số lần thao tác pipet, trong khi câu hỏi nghiên cứu chính của một luận án khoa học máy tính nhiều khi không còn đơn thuần là một thuật toán, mà là một câu hỏi sinh học. Câu hỏi “phân tích dữ liệu này ra sao” giờ xuất hiện ở gần như mọi phòng thí nghiệm, từ nghiên cứu y sinh đến chọn giống vật nuôi, và câu trả lời thường cần đến những người hiểu được cả hai phía.

Tuy vậy, dù bước vào lĩnh vực này từ phía nào, khoảng cách thường lớn hơn người ta tưởng. Một sinh viên sinh học thường có kiến thức sinh học phân tử rất vững, nhưng cùng lúc phải làm quen với dòng lệnh Linux, nhiều thuật ngữ khoa học máy tính xa lạ, các khái niệm thống kê chưa từng học kỹ, vô số công cụ phần mềm, và những bài báo tiếng Anh dày đặc thuật ngữ. Ngược lại, một người đến từ khoa học máy tính, toán hay thống kê có thể viết code và xây dựng mô hình rất tốt, nhưng lại phải học cùng lúc một hệ thống khái niệm sinh học đồ sộ, và quan trọng hơn, phải hiểu dữ liệu được tạo ra trong phòng thí nghiệm như thế nào, mang những sai lệch và giới hạn gì, trước khi có thể đặt ra một câu hỏi có ý nghĩa sinh học. Mỗi mảnh kiến thức đều có tài liệu riêng, nhưng hiếm có tài liệu nào cho thấy các mảnh ấy ghép lại với nhau như thế nào.

Cuốn sách này được viết với mong muốn cung cấp cho bạn đọc đến từ nhiều nền tảng khác nhau, từ sinh học, công nghệ sinh học, y sinh, nông nghiệp đến khoa học máy tính, toán và thống kê, một điểm xuất phát chung để bắt đầu nghiên cứu hệ gen học tính toán. Sách đi qua lịch sử phát triển của hệ gen học, từ những thí nghiệm đầu tiên xác định vật chất di truyền đến khi giải trình tự trở thành một công nghệ thông lượng cao, cách lĩnh vực này dần trở thành một ngành của dữ liệu lớn, và cách các chuyên ngành tính toán như thống kê, thuật toán, học máy và tính toán hiệu năng cao được tích hợp vào sinh học, giúp các nhà khoa học rút ra những hiểu biết mới từ dữ liệu.

Góc nhìn của tác giả

Mọi thứ bắt nguồn từ một sự sắp đặt ngoài dự tính của định mệnh, khi tác giả bắt đầu từ Khoa Sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM. Sau đó, tác giả học thạc sĩ sinh học phân tử thực vật tại Đại học Tel Aviv (Israel), thực tập luận văn tại Khoa Dịch tễ học và Thống kê Y sinh, Viện Karolinska (Thụy Điển), làm tin sinh học tại Viện Nghiên cứu Dữ liệu lớn Vingroup với các dự án hệ gen người Việt, rồi làm tiến sĩ tin sinh học tại Đại học Khoa học Sự sống Na Uy về regulatory genomics trên cá hồi Đại Tây Dương và các loài vật nuôi, song song với thạc sĩ khoa học dữ liệu tại Đại học Texas ở Austin. Hiện tại, tác

giả đang làm việc tại Khoa Y, Đại học Oslo, nghiên cứu di truyền học thống kê của các tính trạng phức tạp và rối loạn tâm thần.

Những công việc đó đưa tác giả qua rất nhiều loại dữ liệu, từ SNP array, short read đến long read, từ transcriptome đến epigenome, và qua nhiều lớp phương pháp, từ phát triển thuật toán, xây dựng pipeline đến di truyền học thống kê và deep learning. Tác giả không viết cuốn sách này từ vị trí của một nhà nghiên cứu lâu năm nhìn lại cả sự nghiệp dài, thay vào đó, tác giả viết từ vị trí của một người vẫn còn nhớ khá rõ cảm giác của người mới bắt đầu, khi phải tự ghép nối kiến thức từ sinh học, khoa học máy tính và thống kê, ba ngành với ba ngôn ngữ, ba cách tư duy và ba cách đặt câu hỏi khác nhau.

Nhìn lại, điều tác giả cần nhất khi mới bắt đầu không phải là một cuốn sách giải thích thật sâu từng thuật toán, hay giải thích quá kỹ về sinh học. Điều tác giả cần lúc đó là một bức tranh tổng quan về những gì người mới bắt đầu sắp phải làm, và một chút động lực để bắt đầu. Đó cũng là điều mà cuốn sách này muốn mang đến cho bạn đọc.

Một chút ngoài lề, cách tiếp cận này cũng chính là cách tác giả đến với khoa học tính toán. Còn nhớ ngày ấy, khi đang ở sa mạc Arava (một vùng hẻo lánh ở miền nam Israel), tác giả tình cờ thấy quảng cáo khóa học về AI Elements of AI¹ của Đại học Helsinki trên fanpage của Đại sứ quán Phần Lan. Với sự tò mò của một sinh viên trẻ, tác giả đăng ký theo học. May mắn là khóa học không quá khó để hoàn thành, nhưng đủ để nhen nhóm một chút đam mê, và không lâu sau đó, tác giả bắt đầu với tin sinh học!

Mục đích và phạm vi

Mục đích của cuốn sách là giúp bạn đọc, dù xuất phát từ sinh học hay từ các ngành tính toán, có kiến thức rộng và cái nhìn toàn cảnh về hệ gen học tính toán, để có thể nhanh chóng bắt đầu tham gia nghiên cứu khoa học. Để làm được điều đó, tác giả chấp nhận một sự đánh đổi có chủ ý, đó là ưu tiên chiều rộng hơn chiều sâu. Nhiều chứng minh toán học, chi tiết kỹ thuật và tham số của từng công cụ đã được lược bớt. Thay vào đó, cuốn sách tập trung vào hai điều. Thứ nhất là những gì bạn sẽ gặp hằng ngày, như làm việc với Linux, giao tiếp với máy tính qua dòng lệnh, sự hoài nghi của một nhà thống kê hay tinh thần tối ưu hóa của một người làm khoa học tính toán. Thứ hai là động lực nghiên cứu, tức là vì sao một phương pháp ra đời, nó giải quyết vấn đề gì, và giới hạn của nó nằm ở đâu. Công cụ phần mềm thay đổi từng năm, nhưng câu hỏi khoa học và nguyên lý phía sau thì bền vững hơn nhiều. Hiểu được động lực, bạn sẽ tự tìm được công cụ phù hợp, kể cả những công cụ chưa ra đời khi cuốn sách này được viết.

Về cách tiếp cận, cuốn sách được tổ chức quanh một ý đơn giản. Mọi thứ đều bắt đầu từ một thực tế không mấy vui vẻ là dữ liệu luôn có nhiễu (noise), và không có công nghệ nào hoàn hảo. Mỗi giới hạn của công nghệ lại sinh ra một bài toán tính toán, và thách thức đó cũng là cơ hội để chúng ta làm tốt hơn. SNP array rẻ nhưng chỉ đo được những vị trí đã chọn trước, nên cần imputation. Short read không nhìn thấy trọn vẹn các vùng lặp và structural variant, nên cần thuật toán và mô hình để suy ra những gì không đo trực tiếp được. Dữ liệu nào cũng có lỗi và nhiễu, nên cần định lượng bất định và kiểm soát false positive. Khi đã hiểu dữ liệu được tạo ra như thế nào và giới hạn của nó nằm ở đâu, lý do phải học thống kê, thuật toán và lập trình sẽ trở nên hiển nhiên. Vì vậy, cuốn sách bắt đầu từ dữ liệu, rồi mới đến các phương pháp.

Về phạm vi, cuốn sách tập trung vào hệ gen học tính toán (*computational genomics*). Các chủ đề được đề cập chủ yếu liên quan đến dữ liệu DNA và RNA từ các công nghệ giải trình tự, các analysis pipeline từ dữ liệu thô đến phân tích ở quy mô quần thể, cùng với hạ tầng tính toán cần thiết để làm việc với dữ liệu lớn. Đây là những chủ đề mà tác giả đã có trải

¹<https://www.elementsofai.com/>

nghiệm nghiên cứu thực sự, hoặc đủ tự tin để trình bày. Những chủ đề như cấu trúc protein, thiết kế thuốc hay các cơ sở dữ liệu protein nằm ngoài phạm vi của sách; bạn đọc quan tâm có thể tìm đến các giáo trình tin sinh học tổng quát.

Về các con số trong sách, những số liệu như kích thước hệ gen, số gen hay tỉ lệ trình tự không mã hóa đều là ước tính, và đã được làm tròn để dễ nhớ thay vì chính xác tuyệt đối. Bản thân các ước tính này cũng thay đổi theo thời gian, khi công nghệ giải trình tự, hệ gen tham chiếu và thuật toán phân tích được cải tiến; số gen mã hóa protein ở người là một ví dụ điển hình, đã được điều chỉnh nhiều lần qua từng giai đoạn nghiên cứu. Đây không phải là khiếm khuyết của sách, mà phản ánh đúng bản chất của khoa học dữ liệu, trong đó mọi con số đều là kết quả của một phép đo và một mô hình, nên luôn có thể được cập nhật. Nói cách khác, đó là một feature, không phải một bug.

Cách đọc cuốn sách

Không có cách đọc nào là đúng duy nhất, vì thứ tự các chương chỉ phản ánh góc nhìn cá nhân của tác giả về một con đường hợp lý để bước vào lĩnh vực này, tức là hiểu dữ liệu trước, rồi đến thống kê, công cụ và các phương pháp phân tích. Bạn thấy phần nào cần cho mình thì cứ đọc trước, phần nào đã quen thì cứ mạnh dạn bỏ qua.

Về thuật ngữ, tác giả giữ nguyên tiếng Anh cho những từ mà giới nghiên cứu dùng hằng ngày, như central dogma, read, alignment hay variant calling. Đơn giản vì sớm muộn gì bạn cũng sẽ đọc bài báo, tài liệu phần mềm và trao đổi với đồng nghiệp quốc tế bằng chính những từ này; Việt hóa gượng ép chỉ thêm một bước “dịch ngược” trong đầu. Những khái niệm đã quen thuộc như hệ gen, phiên mã hay dịch mã thì vẫn giữ tiếng Việt, và cuối mỗi chương có bảng thuật ngữ để tra cứu.

Một lời cuối

Hệ gen học tính toán thay đổi rất nhanh, vì vậy chắc chắn sẽ có nhiều phần của cuốn sách trở nên lỗi thời, và sai sót là điều khó tránh khỏi. Mọi góp ý từ đồng nghiệp và bạn đọc đều được trân trọng, và sẽ giúp những lần tái bản sau tốt hơn. Nếu sau khi đọc xong, bạn có thể mở một bài báo về sinh học tính toán, hiểu được người ta đang làm gì và vì sao họ làm vậy, rồi tự tin bắt tay vào dự án nghiên cứu đầu tiên của mình, thì cuốn sách đã hoàn thành nhiệm vụ.

Cuốn sách này xin được dành để tưởng nhớ cô Lưu Thị Thanh Tú.

Drammen, tháng 09 năm 2026

Dat Nguyen