Chuột chù đuôi bút thích uống rượu dừa. Rượu có khả năng gây nguy hiểm đến tính mạng cho động vật có vú. Các loài chuột chù đã phát triển khả năng tiêu thụ rượu đẻ tránh việc say xỉn. Vì những tác động của rượu lên loài này, các nhà khoa học tin rằng cồn có thể giúp loài chuột chù có các đặc điểm tiến hoá có lợi như chống lại đột quỵ. Tinh tinh cũng có thể là loài “nghiện rượu” vì thích nước hoa quả lên men.

Loài men vi sinh *Saccharomyces cerevisiae* biến đường glucose thành cồn.

Khi nguồn cung glucose cạn kiệt, với tác động của oxy, loài này sẽ thay đổi khả năng tiêu hoá của mình và đổi sang tiêu hoá cồn mà chúng đã sản xuất. Đây được gọi là sự phát triển của diauxic. Nếu không có oxy, loài này sẽ vào trạng thái ngủ đông đến khi có thêm glucose hoặc có oxy. Vì vậy các nhà làm rượu phải đóng kín rượu trong các thùng.

Sự phát triển của diauxic là một quy trình phức tạp ảnh hướng tới biển hiện của nhiều gen. Điều này hẳn phải từ các sự kiện tiến hoá chính của loài này vì chúng không chỉ có khả năng tạo ra cồn, được coi là chất độc giết chết rất nhiều các loại vi sinh và nấm khác cạnh tranh với chúng, mà còn có thể tiêu thụ cồn.

Sự kiện tiến hoá hiếm hoi được gọi là sao chép toàn bộ genome (WGDs). Ví dụ: triệu năm trước, loài cây có quả đầu tiên xuất hiện nhưng không loài nào có thể tiêu thụ được glucose được tạo bởi trái cây. Loài đầu tiên có thể tiêu thụ được đã có được sự tiến hoá vượt bậc. Thay vì tạo ra gen mới, sự phát triển của diauxic đã yêu cầu tạo ra phương thức tiêu thụ mới mà nhiều gen cùng làm việc.

WGD không được ủng hộ nhiều vì chỉ 13% gen của S.cerevisiae được nhân đôi. Wolfe và Shields đã đưa ra tính toán ủng hộ WGD đầu tiên. Việc chỉ có 13% gen được nhân đôi là không quá ngạc nhiên khi kể cả có hàng trăm gen góp phần trông sự tiến hoá cải tiến thì sau WGD, hầu hết các gen không cần cho sự cải tiến. Các gen không cần này sẽ bị phá huỷ bởi các đột biến đến khi chúng biến thành các mã gen giả (không biên dịch, cấu hình đặc tính) và biến mất khỏi bộ gen sau hàng triệu năm.

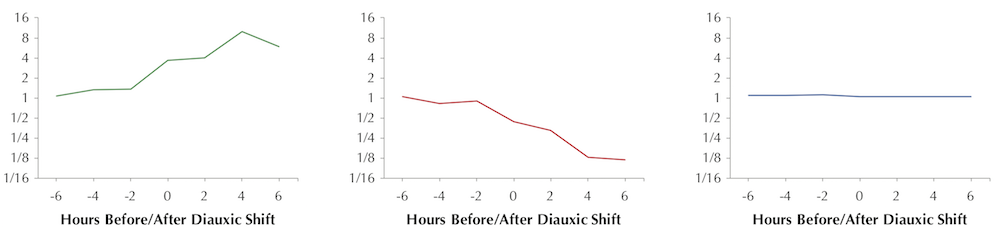
Trong quá trình lên men, nấm men sẽ chuyển đổi acetaldehyd thành cồn, nếu có oxy thì cồn tích luỹ sẽ trở thành acetaldehyd. Cả 2 sự chuyển đổi này đều được xúc tác bởi enzyme alcohol dehydrogenase (Adh). Ở S.cerevisiae, ADH được mã hoá bới Adh1 và Adh2, phát sinh từ cùng 1 gen tổ tiên. Adh1 có khả năng sản xuất cồn còn Adh2 có khả năng tiêu thụ cồn.

Thí nghiệm của Michael Thomson sử dụng nhiều gen Adh từ các loài nấm men để tìm ra gen Saccharomyces cổ đại. Gen này cho thấy tính chất đổi từ acetaldehyd thành cồn như Adh1. Thomson kết luận rằng trước WGD trong loài Saccharomyces, Adh chủ yếu chỉ tạo ra cồn, không tiêu thụ. Sau WGD, Adh1 thực hiện chức năng ban đầu còn Adh2 tiêu thụ cồn để cung cấp năng lượng cho sự phát triển Diauxic.

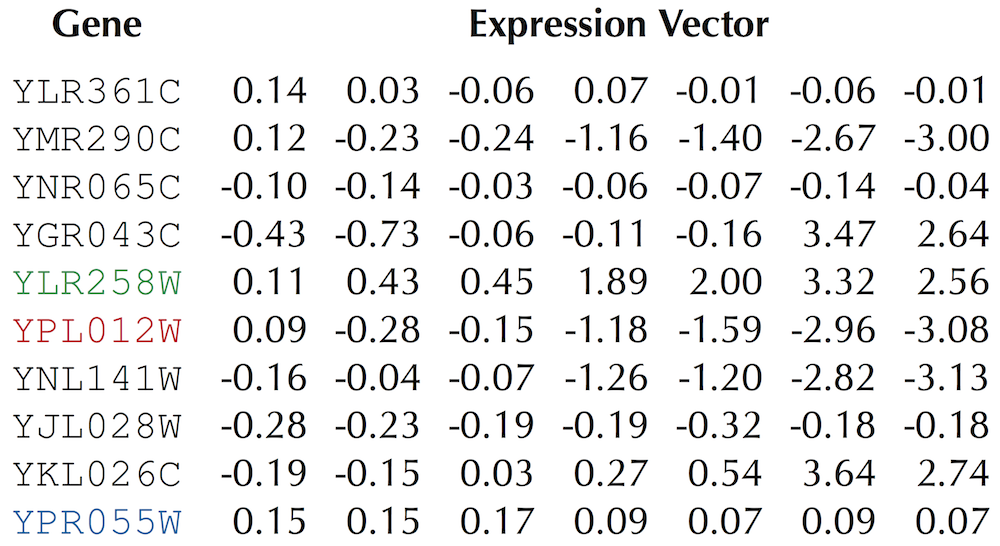
Biểu diễn tất cả n gen nấm men tại m thời điểm ở 2 phía của sự phát triển Diauxic cho ra ma trận đặc tính gen n x m E. Ei,j là giá trị thể hiện cấp độ biểu hiện đặc tính của gen I tại thời điểm j. Cột thứ i được gọi là vecto biểu hiện của gen i.

Joseph DeRisi đã tổng hợp nên ma trận biểu hiện gen đầu tiên có kích thước 6400 x 7 bằng cách thử mẫu mỗi 2 tiếng trong 6 tiếng trước và sau khi sự phát triển của diauxit.

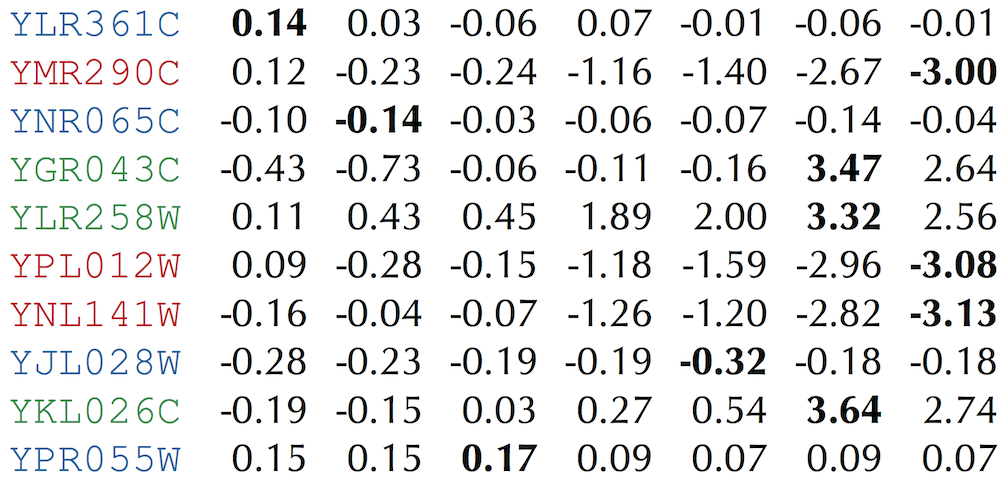
Derisi đã phải dùng phương pháp microarrays.



Gen YPR055W gần như phẳng trong quá trình phát triển của diauxit nên ta kết luận rằng gen này không liên quan đến quá trình này. Gen YLR258W thay đổi đáng kể trong quá trình này nên có thể đây là gen liên quan tới quá trình này. Kiểm tra trong cơ sở dữ liệu gen thì đây chính là gen mã hoá enzyme tổng hợp glucose. Enzyme này điều khiển công việc sản xuất glycogen, một polisacarit glucose dùng để chứa glucose trong tế bào nấm men.



Dù sự phát triển của diauxic là một sự kiện quan trọng trong vòng đời của S.cerevisae, nó không ảnh hưởng hầu hết các chức năng của nấm men. Giả định rằng hầu hết gen của S.cerevisae không biến động trong quá trình phát triển của diauxit, loại trừ các gen này trong quá trình xét, giảm kích thước ma trận. Loại bỏ các gen mà vecto đặc tính trong khoảng −δ và δ (ở đây chọn δ = 2.3) giảm 6400 gen xuống còn 230 gen.



Mục tiêu **phân chia** tập hợp tất cả các gen nấm men thành **k cụm** rời rạc để các gen trong cùng một cụm có các vectơ biểu hiện tương tự. Số lượng cụm không được biết đến trước, và vì vậy thường áp dụng các thuật toán phân cụm cho dữ liệu biểu hiện gen cho các giá trị khác nhau của *k*, chọn giá trị của *k* có ý nghĩa về mặt sinh học.

Để xác định các nhóm gen có kiểu biểu hiện tương tự, chúng ta sẽ nghĩ về một vectơ biểu hiện có chiều dài *m* như một điểm trong không gian *m* chiều; Do đó, các gen có vectơ biểu hiện tương tự sẽ tạo thành các cụm điểm lân cận. Lý tưởng nhất, các cụm phải thỏa mãn nguyên tắc thông thường sau đây, được minh họa trong hình dưới đây cho *m* = 2.

Thay vì chia các điểm dữ liệu *Dữ liệu* thành các cụm *k*, chọn một tập hợp các *Trung tâm* của *k* điểm sẽ đóng vai trò là **trung tâm** của các cụm này, để chúng giảm thiểu một số chức năng khoảng cách giữa *Trung tâm* và *Dữ liệu* trên tất cả các lựa chọn có thể có của trung tâm.

Xác định **khoảng cách Euclid** giữa các điểm *v* = (*v1*, ... , *vm*) và *w*= (*w1*, , , *wm*) trong không gian m chiều, ký hiệu *là d* (*v*, *w*), là độ dài của đoạn thẳng nối các điểm này.



Cho một điểm *DataPoint* trong không gian đa chiều và một tập hợp *k* điểm Centers, chúng ta xác định khoảng cách từ *DataPoint* đến *Centers*, ký hiệu là *d* (*DataPoint*, *Centers*), là khoảng cách Euclid từ *DataPoint* đến tâm gần nhất của nó.

*d*(*DataPoint*,*Centers*) = minall points*x*from*Centersd*(*DataPoint*, *x*).

Độ dài của các phân đoạn tương ứng với *d* (*DataPoint*, *Centers*) cho mỗi điểm *DataPoint*.

Xác định khoảng cách giữa tất cả các điểm dữ liệu Trung tâm *dữ liệu* và *trung tâm*. Khoảng cách này, ký hiệu *là MaxDistance* (*Dữ liệu*, *Trung tâm*), là tối đa của *d* (*DataPoint*, *Centers*) trong số tất cả các điểm dữ liệu *DataPoint*,

*MaxDistance*(*Data*, *Centers*) = maxall points*DataPoint*from*Data d*(*DataPoints*,*Centers*).

