Sequence Alignment اليمنت

فصل سی و یکم از سری کتب الکترونیکی رایگان

پرتال بیوانفورماتیک ایرانیان www.ibp.ir

درج کننده مطلب : بابک باباعباسی

منبع : کارگاه الکترونیکی بیوانفورماتیک ،انجمن بیوتکنولوژی به مدیریت دکتر ملبوبی

مقايسه دو توالي

در دهه ۸۰، يک محقق هيچ برنامه رايانه‌اي براي اين كه بتواند بين تعدادي توالي مشابه توالي را به توالي خود پيدا كند نداشت. بنابراين زبده ترين دانشمندان ان روزگار نيز مجبور بودند اين كار را به صورت دستي انجام دهند. به طور مثال، اگر قرار بود از بين چهار توالي زير مشابه ترين توالي را به توالي خود انتخاب مي كرد. بايد تک تک توالي ها را با توالي الگو مقايسه مي كرد و ميزان شباهت‌ها را در هر مقايسه به دست مي آورد. امروزه به اين همرديفي دوگانه (Pairwise Alignment) مي گويند .

AATTGGCCTTGC

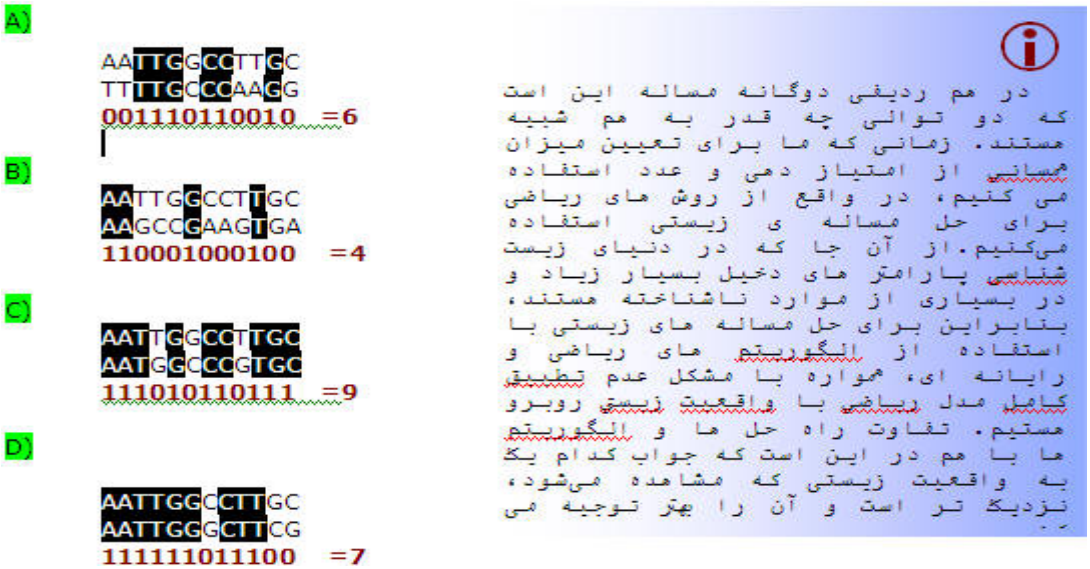
AATGCGCCGTGC

AAGCCGAAGTGA

TTTTGCCCAAGG

ساده ترين راه براي مقايسه كردن دو توالي اين است كه هر بار دو توالي را در زير هم قرار دهيم و يک به يک باز ها را با هم مقايسه كنيم تا شبيه ترين توالي را پيدا كنيم (شكل‌هاي A تا D زير). ولي سوالي كه پيش مي‌آيد اين است كه چگونه و با چه معياري دو توالي مشابه تر را انتخاب مي‌كنيم؟ در اين جاست كه بحث امتياز دهي (Scoring)مطرح مي شود. به عنوان مثال، ساده ترين نوع امتياز دهي اين گونه مي تواند باشد كه اگر دو بازي كه زير هم قرار مي گيرند از يكسان باشند، امتياز ۱ و اگر همسان نباشند، امتياز صفر داده شود. با اين روش مي توان مشابه ترين توالي و درجه شباهت ساير توالي ها را با توالي الگو به دست آورد .

حال توالي‌هاي بالا را با توالي الگوي داده شده با همين روش همرديفي دو گانه مي‌كنيم. به نظر مي‌رسد مورد C كه امتياز ۹ گرفته است، همساني بيشتري با توالي ما دارد.



اما همانطور كه در شكل زير ديده مي‌شود، اين دو توالي را به طريق ديگري هم مي‌توان همرديف كرد .لذا سوالاتي هنوز باقي است مانند آن كه آيا همرديفي ديگري ممكن است؟ كدام همرديفي بهتر است؟ همرديفي بهتر يعني چه؟ كدام همرديفي گوياي اتفاقات زيستي است؟ آيا در همرديفي‌ها روندهاي تكاملي قابل رديابي است؟ تا چه حد؟ و چگونه مي‌توان همرديفي‌ها در اين راستا به كار گرفت؟

اينها سوالات عميقي است كه پايه‌هاي اساسي داده‌پردازي زيستي را تشكيل مي‌دهند. لكن در اين دوره سعي بر آن است كه اصول همرديفي تا اندازه‌اي آموزش داده شود تا به توان از آن براي جستجوي توالي‌هاي مشابه و قضاوت در مورد ميزان مشابهت و درك مفهوم خانواده‌هاي ژني و پروتئيني استفاده نمود .

C)

```
AATTGGCCTTGC
AATGGCCCGTGC
111010110111   =9
```

E)

```
AATTGGCCT-TGC
 AATGGCCCGTGC
101111100111   =9
```

روش Dot plot

به عنوان راهي براي شناسايي تمامي همرديفي‌هاي ممكن محققين، روشي گرافيكي به نام دات پلات (dot plot) به كار بردند. در اين روش دو توالي به صورت عمود بر هم روي محور x ها و y ها در يک صفحه قرار داده مي شوند و در هر نقطه اي كه شبيه هم باشند عدد يك قرار داده مي شود. اگر دو توالي كاملا شبيه باشند

	A	A	T	T	G	G	C	C	T	T	G	C
A	1	1	0	0	0	0	0	0	0	0	0	0
A	1	1	0	0	0	0	0	0	0	0	0	0
T	0	0	1	1	0	0	0	0	1	1	0	0
G	0	0	0	0	1	1	0	0	0	0	1	0
G	0	0	0	0	1	1	0	0	0	0	1	0
C	0	0	0	0	0	0	1	1	0	0	0	1
C	0	0	0	0	0	0	1	1	0	0	0	1
C	0	0	0	0	0	0	1	1	0	0	0	1
G	0	0	0	0	1	1	0	0	0	0	1	0
T	0	0	1	1	0	0	0	0	1	1	0	0
G	0	0	0	0	1	1	0	0	0	0	1	0
C	0	0	0	0	0	0	1	1	0	0	0	1

در نهايت، از رسم نقاط يک خط اوريب بدون شكستگي را مي توان از انتهاي سمت راست صفحه به انتهاي سمت براست بالاي صفحه رسم كرد. همرديفي در واقع مشخص كردن رابطه ي بين نوكلئوتيد هاي يک توالي با توالي ديگر است. اگر دو توالي در مثال فوق را به صورت دات پلات در آوريم جدول زير به دست خواهد آمد. اگر دور بيش از دو عدد ۱ به دنبال هم خط بكشيم، منظره بالا ظاهر خواهد شد. حال بر اساس اين خطوط اريب مي‌توان كليه همرديفي‌هاي دو گانه را استخراج كرده و به صورت خطي نوشت. در اين جدول، دو رابطه خطي C و E در شكل قبل با خطوط آبي و قرمز نشان داده شده‌اند .

مقايسه همرديفي‌ها

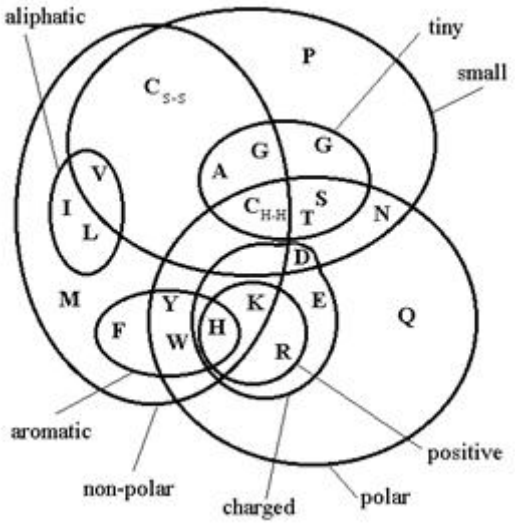
در مثال‌هاي فوق روش امتياز دهي صفر ويك را مي توان يک نوع الگوريتم به حساب آورد كه بر اساس آن همرديفي با بالاترين امتياز را به عنوان بهترين همرديفي انتخاب كرديم. ولي در عمل مي‌توان همرديفي‌هايي را مثال زد كه با وجود امتياز مساوي يا حتي بالاتر صحيح نبوده و با دانسته‌هاي قبلي تطبيق نمي‌كنند. بنابراين تلاش زيادي براي طراحي و بكارگيري الگوريتم‌هايي دقيق‌تر صورت مي‌گيرد كه تا هر چه بيشتر دربرگيرنده واقعيات زيستي و اصول حاكم بر حيات باشد .

به طور مثال، در همرديفي توالي‌هاي نوكلئوتيدي مي‌توان بين امتيازات جايگزيني‌هاي از نوع جانشيني (Substitution) و انتقال (Transition) تفاوت قايل شد. زيرا با توجه به

ساختمان دو رشته‌اي DNAاحتمال جايگزيني بازهاي پوريني با هم و باز هاي پيريميديني با هم بيشتر است. در حالي كه الگوريتم قبلي تفاوتي را بين اين دو حالت قائل نبود. بنابراين جواب هايي هم كه با الگوريتم قبلي به دست آمد كمتر به واقعيت نزديک است .

اين مشكل در توالي‌هاي پروتئيني به طور جدي‌تري مطرح است. در اين توالي‌هاي همه جايگزيني‌ها اشكال ساختاري و عملكردي زيادي ايجاد نمي‌كنند. به عبارت ديگر، برخي اسيد آمينه‌ها خواص فيزيكوشيميايي مشابهي دارند (شكل زير) و مي‌توانند با حداقل تغيير خواص جايگزين يكديگر شوند .

نكته ي ديگر اين است كه در همرديفي هاي بالا فرض برابر بودن طول توالي هاست. در حالي كه در تكامل توالي ها هم پديده ي اضافه شدن را داريم و هم پديده حذف اتفاق مي‌افتد .بنابراين اين انتظار مي رود در بسياري از موارد دو توالي را با هم مقايسه كنيم كه داراي طول يكساني نباشند .



جمع‌بندي اين مقدمات نشان مي‌دهد مي‌توان با كمي‌نمودن (امتيازدهي) نتايج همرديفي آنها را باهم مقايسه نمود. البته براي كمي نمودن همرديفي‌ها حداقل دو نوع امتياز دهي را بايستي منظور كرد .

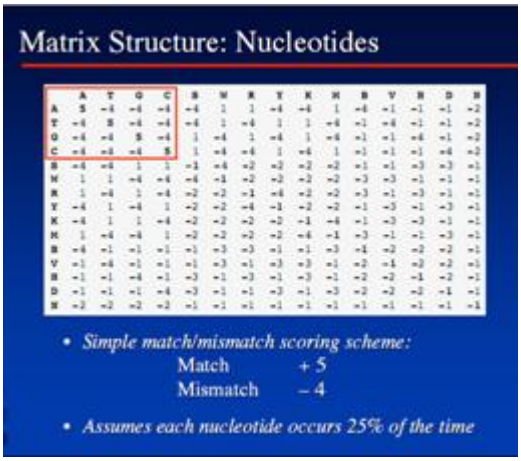
امتياز دهي جايگزيني‌ها

امتياز دهي حذف و اضافه شدن توالي‌ها

به اين ترتيب، امتياز هر همرديفي جمع جبري كليه امتيازانت جايگزيني‌ها و حذف يا اضافه‌ها خواهد بود.

با درك اين كه روش صفر ويك كفايت نمي‌كند و جايگزيني نوكلئوتيدها يا اسيد آمينه‌ها با يكديگر امتياز منفي يا مثبت مساوي ندارند.، متخصصين امر در پي تهيه جداول امتيازدهي جايگزيني‌ها (Subtitution Scoring Matrices) بوده‌اند. به طوري كه تا حد امكان واقعيت‌هاي زيستي را منعكس نمايد .

براي توالي‌هاي نوكلئوتيدي كار چندان دشوار نيست زيرا هر گونه جايگزيني منجر به جهش مي‌شود كه اثر آن در رمزدهي پروتئين‌ها ممكن است مشاهده شود. يعني در اين مولكول‌هاي بحث ساختار و عمل چندان مطرح نيست. البته با توجه به ساختمان دو رشته‌اي DNAمتخصصين تكامل زيستي بين جانشيني نوكلئوتيد پورين و پيريميدين و انتقال از پورين به پيريميدين يا بالعكس تفاوت قائلند. جدول زير نمونه‌اي از جداول امتيازدهي براي همرديفي دو توالي نوكلئوتيدي را نشان مي‌دهد. در اين جدول به طور ساده‌اي كليه همساني‌ها امتياز ۵+ و براي غير جفت شدگي امتياز ۴- در نظر گرفته شده است. ساير حروف در صورت وجود انتخاب براي دو نوكلئوتيد يا بيشتر در هر موقعيت از توالي كاربرد دارند .



در تهيه جداول امتيازات جايگزيني توالي‌هاي پروتئيني خواص فيزيكوشيميايي اسيدهاي آمينه و تاثير جايگزيني آنها در ساختار و عمل پروتئين‌ها مطرح است. در گذشته، پژوهشگران به گروه‌بندي اسيدهاي آمينه بر اساس خواص آنها (شكل صفحه قبل) مراجعه كرده و ميزان مشابهت را به صورت توصيفي (و نه عددي) بيان مي‌كردند. در دو دهه اخير روش‌هاي تهيه جداول امتيازدهي جايگزيني مبتني بر داده‌هاي موجود در طبيعت بوده است. با فرض بر اين كه اگر دو اسيد آمينه داراي خواص فيزيكوشيميايي مشابهي هستند بايستي در طول تكامل جايگزيني آنها تحمل شده باشد، پژوهشگران نسبت به جمع‌اوري توالي‌ها، همردفي آنها با هم و محاسبه فراواني جايگزيني‌ها در بين پروتئين‌هاي همخانواده اقدام نمودند .

در اولين تلاش، با همرديفي ۱۵۷۲ توالي پروتئيني در ۷۱ درخت از ۳۴ خانواده پروتيئني گروه‌بندي شدند. سپس فراواني جايگزيني يك اسيد آمينه با اسيد آمينه ديگر در فرمول زير بكار گرفته شد :

PAM_n= nlog Probability of one substitution/Probablity of occurring by chance*100

در اين فرمول يك واحد PAM (**P**oint **A**ccepted **M**utation) معادل تغيير ۱ در يك توالي صدتايي از اسيدآمينه‌هاست. داده‌هاي حاصل در جدول PAM ثبت مي‌شود. از آنجا كه در طول تكامل ممكن است اسيدآمينه در يك موقعيت چندين بار جايگزين شود، جدول حاصل را چندين بار در خود ضرب مي‌كنند. به طور مثال، براي تهيه جدول PAM_{100}ان را ۱۰۰ بار در خودش ضرب مي‌كنند (جدول a در صفحه بعد .)

بعدها جداول ديگري تدوين شدند كه از يك نوع اصول پيروي مي‌كردند. با اين تفاوت كه فراواني جايگزيني‌ها تنها در مناطق حفاظت شده (Conserved Blocks) براي ساختن جدول وارد محاسبه مي‌شدند. در آن هنگام، ۲۰۰۰ بلوك از ۵۰۰ خانواده پروتئيني در نظر گرفته شد. به طور مشابهي، فرمول زير بكار رفت :

$BLOSUM_{\%}$=log Probability of substitution in block/Probablity of occurring by chance

اين جداول را BLOSUM نامیدن كه از اصطلاح **Blo**ck **Sub**stitution **M**atrix برگرفته شده است. شماره جدول به نوع بلوك مورد استفاده براي محاسبه فراواني و احتمال وقوع جايگزيني بستگي دارد .مثلا BLOSUM62 يعني اين جدول برمبناي فراواني جايگزيني‌ها در بلوك‌هاي حاوي توالي‌هاي با همساني ۶۲ درصد يا بيشتر تشكيل شده است (شكل زير .)

چند اصطلاح:

همسانی یا (Identity) وقتی بکار می‌رود که منظور همردیفی یك نوکلئوتید یا اسیدآمینه با همانند آن منظور باشد.

مشابهت یا (Similarity) وقتی بکار می‌رود که منظور همردیفی یك نوکلئوتید یا اسیدآمینه با معادل آن منظور باشد (شکل دو صفحه قبل و جداول صفحه بعد).

همولوژی (Homology) وقتی بکار می‌رود که منظور همردیفی یك نوکلئوتید یا اسیدآمینه با معادل آن منظور باشد و رابطه نیاکانی بین توالی‌ها در نظر باشد.

(a)

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	5																								G
A	1	2																							A
V	-1	0	4																						V
L	-4	-2	2	6																					L
I	-3	-1	4	2	5																				I
P	0	1	-1	-3	-2	6																			P
S	1	1	-1	-3	-1	1	2																		S
T	0	1	0	-2	0	0	1	3																	T
D	1	0	-2	-4	-2	-1	0	0	4																D
E	0	0	-2	-3	-2	-1	0	0	3	4															E
N	0	0	-2	-3	-2	0	1	0	2	1	2														N
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4													Q
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5												K
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6											R
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6										H
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9									F
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	0	7	10								Y
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	2	-3	0	0	17							W
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6						M
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12					C
B	0	0	-2	-3	-2	-1	0	0	3	3	2	1	1	-1	1	-4	-3	-5	-2	-4	3				B
Z	0	0	-2	-3	-2	0	0	-1	3	3	1	3	0	0	2	-5	-4	-6	-2	-5	2	3			Z
X	-1	0	-1	-1	-1	-1	0	0	-1	-1	0	-1	-1	-1	-1	-2	-2	-4	-1	-3	-1	-1	-1		X
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1	*
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	

(b)

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	6																								G
A	0	4																							A
V	-3	0	4																						V
L	-4	-1	1	4																					L
I	-4	-1	3	2	4																				I
P	-2	-1	-2	-3	-3	7																			P
S	0	1	-2	-2	-2	-1	4																		S
T	-2	0	0	-1	-1	-1	1	5																	T
D	-1	-2	-3	-4	-3	-1	0	-1	6																D
E	-2	-1	-2	-3	-3	-1	0	-1	2	5															E
N	0	-2	-3	-3	-3	-2	1	0	1	0	6														N
Q	-2	-1	-2	-2	-3	-1	0	-1	0	2	0	5													Q
K	-2	-1	-2	-2	-3	-1	0	-1	-1	1	0	1	5												K
R	-2	-1	-3	-2	-3	-2	-1	-1	-2	0	0	1	2	5											R
H	-2	-2	-3	-3	-3	-2	-1	-2	-1	0	1	0	-1	0	8										H
F	-3	-2	-1	0	0	-4	-2	-2	-3	-3	-3	-3	-3	-3	-1	6									F
Y	-3	-2	-1	-1	-1	-3	-2	-2	-3	-2	-2	-1	-2	-2	2	3	7								Y
W	-2	-3	-3	-2	-3	-4	-3	-2	-4	-3	-4	-2	-3	-3	-2	1	2	11							W
M	-3	-1	1	2	1	-2	-1	-1	-3	-2	-2	0	-1	-1	-2	0	-1	-1	5						M
C	-3	0	-1	-1	-1	-3	-1	-1	-3	-4	-3	-3	-3	-3	-3	-2	-2	-2	-1	9					C
B	-1	-2	-3	-4	-3	-2	0	-1	4	1	3	0	0	-1	0	-3	-3	-4	-3	-3	4				B
Z	-2	-1	-2	-3	-3	-1	0	-1	1	4	0	3	1	0	0	-3	-2	-3	-1	-3	1	4			Z
X	-1	0	-1	-1	-1	-2	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	-2	-1	-1	-1		X
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1	*
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	

منابع زیادی در تهیه این کتاب الکترونیک استفاده شده است که در بخش معرفی کتب پرتال بیوانفورماتیک ایرانیان میتوانید مطالعه کنید www.ibp.ir

منبع اصلی این مطلب کارگاه بیوانفورماتیک به مدیریت دکتر ملبوبی میباشد

****
****

دوست ارجمند این مطلب آموزشی به منظور آشنایی هر چه بیشتر ایرانیان با علم بیوانفورماتیک تهیه شده است خواهشمند است از هر گونه سو استفاده از این مطالب بپرهیزید زیرا که این مطالب تحت حمایت قانون کپی رایت میباشند

این مطالب به این دلیل به صورت کتابهای الکترونیکی تهیه شده اند تا
هرکس، در هرجا ، در هر زمان و به صورت رایگان بتواند به این مطالب دست رسی داشته باشد پس اگر چنانچه تمایل به استفاده از این مطالب در سایت یا وبلاگ خود را دارید این کار با ذکر منبع بلامانع میباشد.

هیچ شخص حقیقی یا حقوقی حق ندارد تا از مطالب کتب الکترونیکی این پرتال به منظور چاپ و نشر کتاب استفاده کند چاپ و نشر کتاب کاری مقدس میباشد ولی فروش علم جزء ناشایسته ترین کارهاست که متاسفانه برخی با استفاده از نا آشنایی مردم با منبع سرشار اینترنت دست به چاپ کتب میزنند و آن را به قیمت بسیار بالایی به فروش میرسانند
حرف من این است اگر عاشق علم هستید دانش خود را به صورت رایگان در اختیار همگان قرار دهید و بهترین راه آن توسط اینترنت میباشد.
به امید روزی که هیچ کتابی چاپ نشود و تمام مطالب به صورت رایگان در اینترنت قابل دست رسی باشند تا حد اقل در بخش استفاده از منابع علمی نامی از غنی و فقیر برده نشود .

****
****

پرتال بیوانفورماتیک ایرانیان www.ibp.ir
بابک باباعباسی