به نام خدا

محمدرضا منعمیان

شماره دانشجویی : ۴۰۲۱۰۶۶۰۴

Mohammad Reza Monemian

bioinformatic

سوال اول)

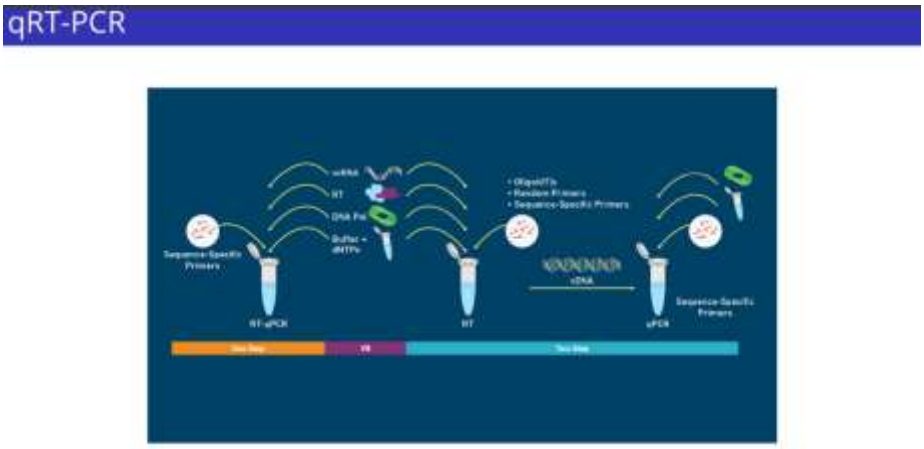
قسمت آ :

ابتدا به بیان دو روش می‌پردازیم :

روش qpcr : در این روش DNA ورودی با استفاده از DNA POLYMERISE و PRIMER هایی که وارد می‌کنیم تکثیر می‌شود و ما حضور یا عدم حضور توالی که primer آن را وارد کردیم در ژنوم خود بررسی کنیم. توجه داریم که DNA همه سلول های بدن یک جاندار مانند هم است و در سلول های مختلف بیان ژن است که متفاوت است بنابراین این روش نمی‌تواند اینکه چه ژنی در ژنوم ورودی بیان شده است را بررسی کند.

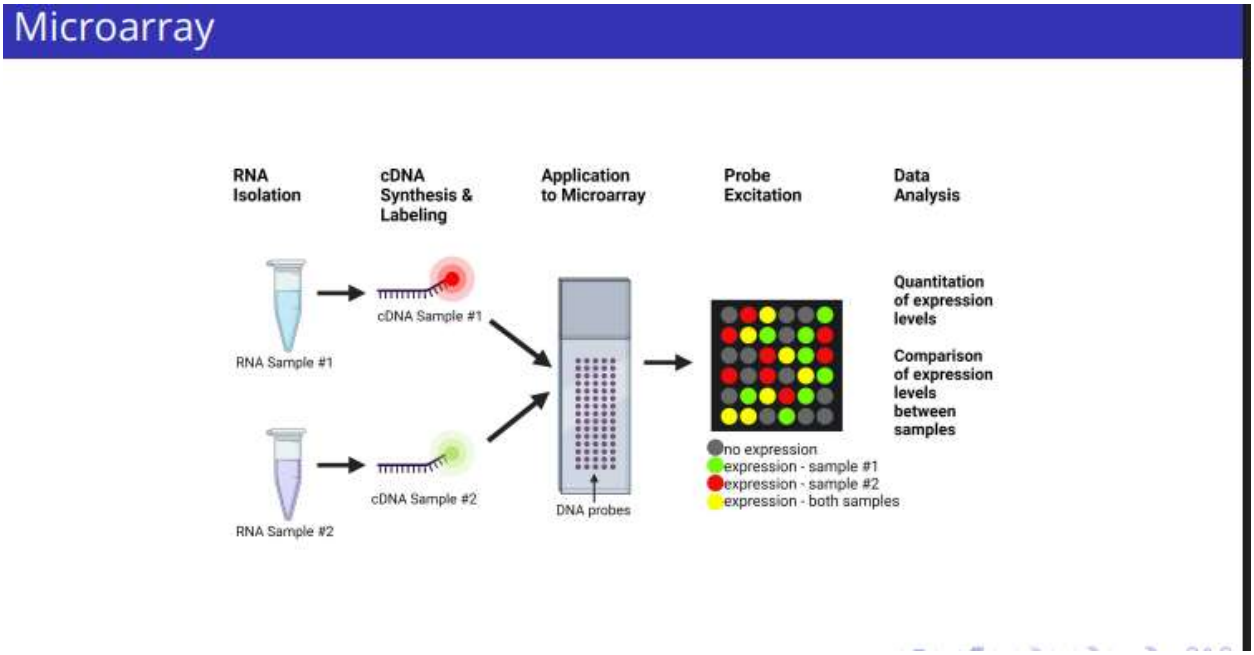
روش qpcr-rt : در این روش ورودی ما به جای DNA ، mRNA است. آنزیمی به نام Reverse transcription وجود دارد که می‌تواند معکوس عملیات ترجمه را انجام دهد یعنی mRNA را به cDNA تبدیل کند. با استفاده از این آنزیم می‌توان ابتدا mRNA ورودی را به cDNA تبدیل کرد و با تکثیر آن به روش قبلی مستقیما بیان ژن را بررسی کرد.

بنابراین به طور خلاصه روش توالی‌یابی qRT-PC برای اندازه‌گیری بیان ژن، از qPCR متمایز است زیرا شامل یک مرحله حیاتی به نام **رونوشت‌برداری معکوس** است که در ابتدای فرآیند انجام می‌شود. از آنجایی که PCR فقط می‌تواند رشته‌های DNA را تکثیر و اندازه‌گیری کند، مرحله RT با استفاده از آنزیم **رونوشت‌بردار معکوس**، اطلاعات RNA(که نشان‌دهنده بیان ژن است) را به یک رشته **DNA** مکمل تبدیل می‌کند. این cDNA سپس به عنوان الگو در واکنش qPCR استفاده می‌شود و به این ترتیب، اندازه‌گیری کمّی محصول نهایی به طور معکوس میزان **RNA اولیه** و در نتیجه سطح **بیان ژن** را در نمونه نشان می‌دهد، که هدف اصلی در مطالعات بیان ژن است.



قسمت ب :

در روش micro array برای مثال ۲ ژنومی که می‌خواهیم بررسی کنیم ابتدا لیبل گذاری می‌شود. سپس آنها بر روی میکروشیپی که دارای پراب های متعدد است قرار می‌گیرد. هر کدام از این پراب ها مکمل یک ژن خاص هستند که هدف مطالعه ماست . وقتی ژنومی مکمل آن پراب باشد به آن نقطه متصل می‌شود. در ادامه هر چقدر نور ساطع شده بیشتر باشد یعنی مولکول DNA بیشتری به آن پراب متصل شده است و این یعنی شدت سیگنال هر نقطه نشان دهنده میزان فراوانی RNA آن ژن در نمونه اصلی است. اما چرا امروزه از روش های NGS بیشتر استفاده می‌شود:



۱.توانایی اکتشافی :

روش micro array محدود به ژن هایی است که پروب های آن از قبل به میکروچیپ متصل شده باشد. اگر در نمونه ما یک جهش یا ژن جدیدی بیان شده باشد این روش قادر به پیدا کردن آن نیست. اما در مقابل روش های جدید می‌تواند کل ترانسکریپتوم را بخواند به همین دلیل می‌تواند به طور همزمان ژن های جدید و ادغام های ژنی و ایزوفروم های پیرایشی را بررسی کند.

۲. دقت و محدوده دینامیکی:

همانطور که اشاره شد روش microarray بیان ژن ها را بر اساس شدت فلوئروسنت می‌سنجد اما به دلیل وجود نویز و یا اشباع سیگنال این روش دارای محدوده دینامیکی کمی است و این باعث می‌شود این روش نتواند دقت بالایی برای بیان های کم محدوده دینامیکی وسیع و مناسبی پیدا کند اما در مقابل روش های جدید با

استفاده از شمارش مطلق تعداد دفعاتی که توالی آن ژن خوانده می‌شود صورت می‌گیرد و این امر سبب ایجاد دقت بالا حتی برای بیان های کم می‌شود.

۳. وارایانت های ساختاری :

روش microarray تنها میزان فراوانی را اندازه می‌گیرد اما روش های NGS می‌تواند ساختار ها و وارایانت های ژنی را نیز تشخیص دهد.

۴. هزینه نهایی روش های NGS بسیار به صرفه تر است.

قسمت ج:

تولید کلاستر فرآیندی است که در آن، هر مولکول DNA الگو روی flow cell تکثیر می‌شود تا **صدها هزار کپی** از یک توالی واحد در یک نقطه مشخص جمع شوند. در اصل ایجاد کلاستر در روش illumine به این دلیل ضروری است که سیگنال نور ساطع شده تقویت شود.
در روش توالی‌یابی از طریق سنتز در هر چرخه، فقط یک نوکلئوتید با یک برچسب فلوئورسنت به رشته‌های DNA در حال رشد متصل می‌شود. سیگنال ساطع شده از یک مولکول بسیار ضعیف است و توسط دوربین های دستگاه به درستی قابل تشخیص نیست. اما وقتی میلیون‌ها کپی یکسان از توالی در یک **کلاستر** جمع می‌شوند، سیگنال‌های نوری آن‌ها با هم جمع شده و به یک **سیگنال نوری قوی** تبدیل می‌شوند که دوربین با وضوح بالا می‌تواند آن را به راحتی ثبت کند.

از طرف دیگر هر کلاستر توالی یک قطعه DNA الگو را به طور مستقل نشان می‌دهد. ایجاد میلیون‌ها کلاستر روی یک سطح جریان‌خانه، دستگاه می‌تواند میلیون‌ها توالی متفاوت را به طور همزمان و موازی بخواند. این ویژگی باعث شده است که illumine به یک روش توالی یابی نسل جدید بدل شود.

قسمت د :

خیر لزوما به این معنی نیست و می‌تواند به دلیل دلایل فنی و بیولوژیکی متعددی باشد که به سطح بیان واقعی RNA ارتباطی نداشته باشد برای مثال :

۱.محدودیت های ساخت و طراحی پروب : اگر توالی پروب DNA مربوط به آن ژن درست طراحی نشده باشد مثال طول نامناسب یا ساختار ثانویه پیچیده کارایی به شدت کاهش می‌یابد و این باعث می‌شود cDNA کمتری به آن متصل شود و سیگنال ضعیف تری ساطع کند.

همچنین اگر در توالی ژن مورد آزمایش جهش یا وارایانتی اتفاق افتاده باشد بطوریکه با توالی پروب طراحی شده متفاوت باشد اتصال به درستی رخ نمی‌دهد و این امر باعث کاهش شدید سیگنال می‌شود حتی اگر بیان ژن بالا باشد.

۲. مشکلات مربوط به آماده سازی :

اگر RNA آن ژن خاص در مراحل اولیه استخراج یا آماده‌سازی (قبل از سنتز DNA)تخریب شده باشد، مواد اولیه کمی برای برچسب‌گذاری فلوئورسنت در دسترس خواهد بود، در نتیجه سیگنال ضعیف می‌شود. همچنین کارایی برچسب فلوئروسنت نیز در مرحله سنتز می‌تواند کاهش یابد و در نتیجه با وجود بیان کافی سیگنال ساطع شده ضعیف باشد.

۳. نویز :

اگر سطح بیان یک ژن به صورت طبیعی بسیار پایین باشد، سیگنال آن ممکن است به سختی از **نویز پس‌زمینه** تفکیک شود و به اشتباه به عنوان بیان کم شناخته شود.

سوال دوم)

قسمت الف :

در روش global عادی ما به دنبال یک assignment به صورت سراسری هستیم بطوریکه امتیاز منفی که برای گپ در نظر می‌گیریم ثابت است اما در حالت affine یک امتیاز منفی زیاد برای باز شدن گپ در نظر گرفته و امتیاز کمتری را به ادامه دادن گپ نسبت می‌دهیم.

برای حل این مسئله سه ماتریس upper , lower , middle تعریف می‌کنیم به صورتیکه در ماتریس middle تنها می‌توان به صورت قطری حرکت کرد. در ماتریس upper به صورت اقفی و در ماتریس lower تنها می‌توان به صورت عمودی رفت. برای پر کردن سه ماتریس گفته شده ابتدا به بیان چگونگی مقدار دهی اولیه آنها می‌پردزایم :

برای ماتریس middle :

$$M_{00} = 0, M_{i0} = -\infty, M_{0j} = -\infty$$

برای ماتریس upper :

$$U_{\mathrm{i}0} = -\infty, U_{0\mathrm{j}} = gapopen + (\mathrm{j} - 1)gapextend\ ,$$

برای ماتریس lower :

$$l_{0j} = -\infty, L_{i0} = gapopen + (i-1)gapextend$$

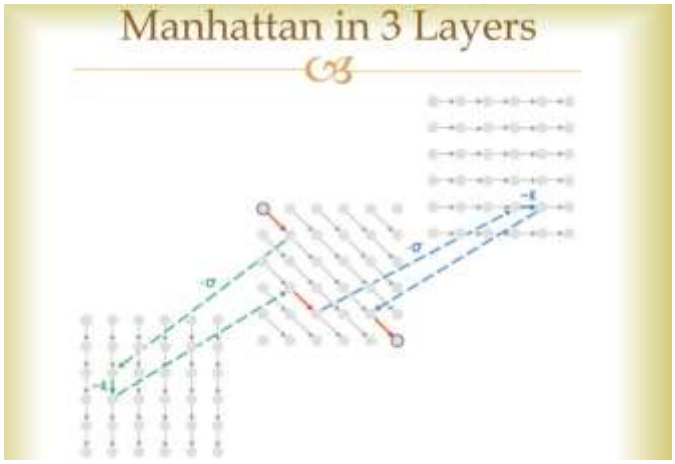
روابط بازشگتی نیز به صورت زیر خواهد بود :

$$L[i][j] = max(L[i-1][j] + gapextend, \mathrm{M}[i-1][j] + gapopen)$$

$$U[i][j] = max(U[i][j-1] + gapextend, M[i][j-1] + gapopen)$$

$$M[i][j] = max(M[i-1][j-1], L[i\text{-}1][j\text{-}1], U[i\text{-}1][j\text{-}1]\,) + \ score(v_i, w_j)$$

که در اینجا منظور از score این است که اگر match رخ داد امتیاز آن و اگر mismatch رخ داد امتیاز مربوط به آن لحاظ شود.



- σ : Jumping penalty is assigned to moving from the main level to either the upper level or the lower level
- ε : Gap extension penalty for each continuation on a level other than the main level)

$$lower_{i,j} = \max \begin{cases} lower_{i-1,j} - \epsilon \\ middle_{i-1,j} - \sigma \end{cases}$$

$$middle_{i,j} = \max \begin{cases} lower_{i,j} \\ middle_{i-1,j-1} + Score(v_i, w_j) \\ upper_{i,j} \end{cases}$$

$$upper_{i,j} = \max \begin{cases} upper_{i,j-1} - \epsilon \\ middle_{i,j-1} - \sigma \end{cases}$$

قسمت ب :

در این قسمت به هم ترازی دو رشته داده شده با استفاده از الگوریتمی که بیان کردیم می‌پردازیم. برای این کار سه ماتریس مورد نظر را پر می‌کنیم :

ماتریس upper :

		T	C	T	C
	*	*	*	*	*
A	-4	*	*	*	*
T	-5	-7 , m	-11,m	-12,m	-13,m
T	-6	-3,m	-10,m	-6,m	-14,u
C	-7	-4,u	-6,m	-2,m	-9,m

ماتریس middle :

		T	C	T	C
	0	*	*	*	*
A	*	-3,m	-7,l	-8,l	-9,l
T	*	1,u	-6,m	-2,m	-11,m
T	*	0,u	-2,m	2,l	-5,m
C	*	-9.u	5,m	-5,m	7,m

ماتریس lower :

		T	C	T	C
	*	-4	-5	-6	-7
A	*	*	-7,m	-8,l	-9,l
T	*	*	-3,m	-4,l	-5,l
T	*	*	-4,m	-5,l	-2,m
C	*	*	-13,m	1,m	0,l

توجه داریم که منظور از * منفی بی نهایت است. همچنین حرفی که کنار امتیاز هر خانه نوشته شده است نشان دهنده جهت حرکت در backtrack است.

همانطور که مشاهده می‌کنیم امتیاز خانه آخر جدول middle برابر با ۷ است و match دو C در آن است. طبق حرف نوشته شده باید قطری حرکت کرد که به خانه ۳و۳ ماتریس middle می‌رسیم. که در آن match دو حرف T را داریم. طبق حرف نوشته شده در این خانه باید به خانه ۲و۲ ماتریس lower برویم و این یعنی در رشته در محور عمودی ماتریس یک گپ ایجاد شده است. اما چون حرف نوشته شده در این خانه m هست این گپ ادامه پیدا نمی‌کند و باید به خانه ۲و۱ از ماتریس middle برویم که در آنجا یک match برای حرف T رخ می‌دهد. در ادامه با توجه به حرف u باید به خانه ۰و۱ ماتریس upper حرکت کنیم و این یعنی در رشته ای که در حالت افقی ماتریس ها قرار گرفته است باید یم گپ ایجاد شود. بنابراین هم ترازی نهایی به صورت زیر است

$-TCTC$
AT-TC

که اگر امتیاز نهایی آن را نیز محاسبه کنیم برابر با ۷ خواهد بود.

سوال سوم)

منظور از قیچی i,j در رشته X,Y بدین معناست که در هم ترازی که به عنوان جواب خروجی داده می‌شود حتما دو کاراکتر x_i, y_j روبه روی هم قرار بگیرند. برای حل این مسئله دو بار باید از الگوریتم نیدلمن وانش استفاده کنیم یکی برای محاسبه امتیاز پیشوند و یکی برای محاسبه امتیاز پسوند. جدول F را برای پیشوند و جدول E را برای پسوند در نظر می‌گیریم. روابط بازگشتی و مقدار دهی اولیه آنها به صورت زیر است:

جدول F :

خانه ۰و۰ آن برابر با صفر است و دیگر درایه های سطر اول و ستون اول آن به صورت زیر محاسبه می‌شود :

$$F[i][0] = (i-1)costofgap, F[0][j] = (j-1)costofgap$$

بقیه سطر و ستونها به صورت بازگشتی به روش زیر محاسبه می‌شوند :

$$F[i][j] \ = \ max(F[i-1][j-1] \ + \ score(x_i, y_j), F[i-1][j] \\ + \ costofgap, F[i][j-1] \ + \ costofgap$$

این جدول برای هر I,j بهترین امتیاز پیشوند را محاسبه می‌کند.

جدول دوم یعنی E برای محاسبه بهترین امتیاز پسوند است. مقدار دهی اولیه آن به صورت زیر و یک طور به شکل مکمل جد.ول F است :

$$\mathrm{E}[i][n] = (m-i)costofgap, G[m][j] = (n-j)costofgap$$

به عبارت دیگر سطر آخر و ستون آخر به ترتیب با حرکت به طرف چپ و حرکت به طرف بالا باید به صورت تساعدی با هزینه گپ پر شود.

برای پر کردن بقیه جدول باید از رابطه بازگشتی زیر استفاده کنیم:

$$\mathrm{E}[i][j] \ = \ max(\mathrm{E}[i+1][j+1] \ + \ score(x_i, y_j), \mathrm{E}[i+1][j] \\ + \ costofgap, \mathrm{E}[i][j+1] \ + \ costofgap$$

حالا با استفاده از این دو جدول با فرمول زیر جدول G را پر می‌کنیم :

$$G[i][j] = F[i\text{-}1][j\text{-}1] + score(x_i, y_j) + E[i+1][j+1]$$

منظور از score نیز یعنی اگر مچ شد امتیاز مچ و اگر نشد امتیاز dismatch محاسبه شود.

واضح است که برای انجام این الگوریتم کافی است دو بار دو حلقه تو در تو به اندازه طول رشته اول و دوم بزنیم یعنی زمان اجرای آن از اردر $|x||y|$ خواهد شد.