شناسایی زنجیره های DNA با استفاده از فیلترهای همسان متغیر با زمان

سعیده سادات بدیعیان
دانشگاه فردوسی مشهد
s_badiyan@yahoo.com

معصومه صفایی
دانشگاه فردوسی مشهد
masoomehsafaie@yahoo.com

چکیده: زنجیره DNA (deoxyribonucleic acid) شامل اطلاعات ژنتیکی هر فرد می باشد. به استخراج این اطلاعات توالی یابی DNA (DNA sequencing) می گویند. از آنجا که خطاهای کوچک در فرایند استخراج اطلاعات سبب ایجاد اشتباهات بزرگ در آزمایشات ژنتیکی می شود، دقت در طی فرایند توالی یابی برای کاهش خطاها بسیار حایز اهمیت می باشد. ما در این مقاله با استفاده از فیلترهای همسان متغیر با زمان به آشکار سازی توالی یابی DNA پرداخته و عوامل موثر در کاهش احتمال خطاها را بهبود بخشیده ایم.

کلمات کلیدی: *DNA، توالی یابیDNA (DNA sequencing)، فیلتر همسان متغیر با زمان (time variant matched filter).*

۱. مقدمه

همانطور که در شکل ۲ دیده می شود DNA به طور معمول از دو رشته(double strands) به هم پیچیده شده تشکیل شده است. هر یک از این رشته های تکی (single strand) از ۴ نوع باز متفاوت تشکیل شده است که بر اساس نوع اطلاعات ژنتیکی، ترتیب و تعداد آنها در رشته متفاوت است. این چهار نوع باز آدنین(A)، گوانین (G)، سیتوزین (C) و تیمین (T) می باشند. یک double strand از اتصال دو single strand که مکمل یکدیگرند تشکیل شده است. A و T مکمل هم و C وG مکمل یکدیگرند.

توالی یابی DNA، یکی از ارکان انقلاب بیوتکنولوژی کنونی است. با فرایند توالی یابی ۲۴ کروموزوم انسانی شناخته شده است. توالی DNA هر آنچه که یک سلول انجام می دهد، از لحظه ای که متولد می شود تا لحظه ای که می میرد، شامل می شود. تغییرات در توالی DNA می تواند شانس انسان را برای مبتلا شدن به یک بیماری افزایش دهد. این تغییرات سبب عدم مقابله بدن در برابر بیماریها یا مقاومت بدن در برابر دریافت داروها می شود.

در بیولوژی از روشهای مختلفی برای توالی یابی استفاده می شود که همگی آنان شامل چند مرحله از جمله PCR (واکنش زنجیره ای پلیمراز) است. در فرایند PCR، DNA به صورت نمایی زیاد می شود. مرحله بعدی توالی یابی به وسیله تئوری الکتروفورسیز توضیح داده می شود که در آن چاهکهای موجود بر روی ژل به عنوان غربالهایی برای ملکول ها عمل می کند به طوری که حرکت انواع مختلف اسید نوکلئیک به طول آنها بستگی دارد. به این ترتیب بر اساس تفاوت در اندازه طول، DNA ها از هم جدا می شوند. از فلورسنت رنگی برای علامتگذاری نوکلئوتیدها استفاده می کنیم که همان طور که در شکل ۳ مشاهده می شود با لیزر آشکار سازی می شوند.

معمولاً سه نوع عمده خطا در هنگام توالی یابی وجود دارد. خطای جابجایی (substitution) زمانی اتفاق می افتد که نوع باز اشتباه تشخیص داده شود. لاگذاری (insertion)، زمانی روی می دهد که در اثر اشتباه یک باز را اضافه خوانده شود. حذف (deletion)، زمانی رخ می دهدکه در اثر اشتباه یک باز خوانده نمی شود. در این مقاله کلمه خطا هر سه نوع را در بر می گیرد. ALF sequencer دستگاه توالی یابی است که تمام خطاهای فوق را شامل می شود.

مخابرات داده بر پایه انتقال و دریافت اطلاعات با احتمال خطای کم استوار است. شکل ۱ یک مدل از سیستم مخابرات داده را نشان می دهد. در ابتدا، ورودی وارد فرستنده شده و یک سیگنال آنالوگ در ورودی کانال حاصل می شود. عبور از کانال دو اثر مهم دارد. ابتدا، شکل موج بوسیله پاسخ ضربه کانال خراب می شود. در مرحله بعد نویز به اطلاعات اضافه می شود. گیرنده، خطاهای آشکار سازی را بوسیله تخمین زدن، حذف ISI و میانگین گرفتن در زمان برای کاهش تاثیر نویز، کاهش می دهد.

در این مقاله، توالی یابی DNA به صورت وابسته به نویز و روی هم افتادگی سیگنال که یک توالی اطلاعاتی را بیان می کند، نشان داده شده است. مخابرات داده نیز برای استخراج یک توالی اطلاعاتی از حالت نویزی و روی هم افتادگی سیگنالها به کار رفته است. به طور واضح یک شباهت قوی بین مخابرات داده و توالی یابی DNA وجود دارد.

در بخشهای بعدی ابتدا یک مدل از سری زمانی DNA ارائه می شود سپس سیستم مخابراتی استفاده شده ارائه می شود و نهایتاً نتایج شبیه سازی مورد بررسی قرار خواهد گرفت.

۲. سری زمانی DNA

دراین بخش یک مدل آماری از سری زمانی DNA بدست آمده از دستگاه توالی یاب معرفی می شود. شکل پیک و پارامترها مدل شده است و نویز کل سیستم یک نویز گوسی جمع شونده سفید فرض می شود.

به طور کلی در سرتاسر ناحیه مرکزی در شکل حاصل از دستگاه توالی یاب (شکل ۴)، یک گرایش رو به پایین در دامنه پیک ها دیده می شود. با قدرت تفکیک بیشتر برای چهار باز متفاوت، مدل سری زمانی به صورت زیر پیشنهاد می شود[1]:

$$y_{n,k} = \sum_{i=1}^{N_b} a_i g_{k,t_i} \delta_{n,x_i} + n_k \qquad (۱)$$

که n نوع باز را نشان می دهد. اندیس k شماره نمونه، جمع روی موقعیت توالی باز i و N_b تعداد باز در توالی است. شکل کلی پالس، g_{k,t_i} است. پیک پالس در t_i واقع شده و با a_i مقیاس بندی شده است. متغیرهای t_i و a_i لرزش زمانی و نوسان دامنه را مدل می کنند. n_k نویز جمع شونده است که نوسان سطح زمینه را معرفی می کند و در حقیقت یک نویز شیمیایی است که به وسیله هر مولکولی که بخش درستی از یک توالی نباشد یا به وسیله آشکارسازها عبور داده شود و یا فلورسنت شده باشد به وجود می آید.

الکتروفورسیز یک مولکول خالص وقتی که مولکولهای DNA به اندازه کافی دور از زمان بارگذاری حرکت کند، باید منجر به یک پیک گوسی شکل شود. اگر چه در نتایج مشاهده شده از دستگاه توالی یاب، شکل پیک بسیار پیچیده تر از گوسی است. در شکل۵ چهار نوع پیک نمونه نشان داده شده است . در شکل ۷خط پیک هر ناحیه حرکت کرده تا پیک دور از مبدا با پیک نزدیک مبدا تطبیق یابد و بنابر این زمان طوری مدرج شده است که بر پیک نزدیک مبدا منطبق شود و پیکها به ارتفاع واحد مدرج شده اند. با توجه به شکل ۷ در می یابیم که پیکها بسیار شبیه یکدیگرند. تنها تفاوت بارز آنها بسط یافتن پیکهای انتهایی در زمان است. پیکها می توانند به صورت فرمول(۲) مدل شوند [1]:

$$g_{k,t_t} = g_1\left(\frac{k - t_i}{p_w(t_i)}\right) \qquad (۲)$$

k، اندیس نمونه، t_i، زمان پیک، g_1، شکل پالس (پیوسته) وقتی عرض پیک واحد باشد و $p_w(t_i)$، نشان دهنده وابستگی عرض پیک به زمان پیک است و به صورت زیر مدل می شود[1]:

$$p_w(t_i) = 15.08 + 0.0326 * (t_i - 1) \quad (۳)$$

در شکل۶، شکل پالس به سه ناحیه مجزا تقسیم می شود که لوب اصلی گوسی و دو دامنه نمایی است.شکل پالس کلی با عرض واحد برای مجموعه داده توالی یابی، فرمول (۴) را نتیجه می دهد [1]:

$$g_1(l) = \begin{cases} 7.89e^{4.8l} & l < -0.86 \\ e^{-(1.67l)^2} & -0.86 < l < 0.64 \\ 1.119e^{-1.96l} & 0.64 < l \end{cases} \quad (۴)$$

برای مدل کردن و شبیه سازی ساختار DNA احتیاج به استفاده از یک سیستم مخابراتی داریم. این سیستم مخابراتی در بخش ۳ توضیح داده می شود.

۳. سیستم مخابراتی استفاده شده

در فرستنده مشاهده شده در شکل ۸ به هر باز یک کد اختصاص می دهیم و بازها ی DNA را با این کدها می شناسیم. A با 00، C با 01، G با 10 و T با 11 نشان داده می شود.

در محل هر باز یک پیک رسم می شود که تنها تفاوت پیک ها در عرض پالس و دامنه است.کانال با نویز سفید گوسی جمع شونده فرض شده است. در گیرنده از همبستیگر استفاده شده است. همبستیگر استفاده شده متغیر با زمان است در هر بازه زمانی خروجی هبستیگر با سطح آستانه(که آن نیز متغیر با زمان است) مقایسه می شود. اگر حاصل بزرگتر از سطح آستانه باشد وجود باز درآن محل با یک پیک نشان داده خواهد شد و در غیر این صورت صفر خواهد بود.(شکلهای ۹و ۱۰)

برای اینکه احتمال خطا کمینه گردد، به روش زیر سطح آستانه را محاسبه می نماییم . اگر خروجی فیلتر همسان به صورت فرمول (۵) می باشد[2]:

$$r(t) = \begin{cases} \sqrt{E} + w & m_0 = 1 \\ w & m_0 = 0 \end{cases} \quad (۵)$$

بطوریکه E انرژی سیگنال و w نویز سفید گوسی جمع شونده با پراش $N_0/2$ می باشد. وجود پیک با یک و عدم وجود پیک با صفر نشان داده می شود، در این صورت تابع چگالی احتمال متغیر پیشای r بصورت زیر است [2]:

$$\begin{cases} f(r/m=1) = \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{(r-\sqrt{E})^2}{N_0}\right) \\ f(r/m=0) = \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{r^2}{N_0}\right) \end{cases} \quad (۶)$$

هر گاه p_1 احتمال ارسال $m_0 = 1$ و p_1 احتمال ارسال $m_0 = 0$ باشد، احتمال خطا برابر است با [2]:

$$p_e = \frac{p_1}{2} erfc\left(\frac{\sqrt{E} - \lambda}{\sqrt{N_0}}\right) + \frac{p_0}{2} erfc\left(\frac{\lambda}{\sqrt{N_0}}\right) \quad (۷)$$

بطوریکه λ سطح آستانه می باشدو سطح آستانه کمینه برابر است با [2]:

$$\lambda_{\min} = \frac{N_0}{2\sqrt{E}}\left(\ln\frac{P_0}{P_1} + \frac{E}{N_0}\right) \qquad (۸)$$

اگر در رابطه ۸، $p_0 = p_1 = \frac{1}{2}$ باشد: $\lambda_{opt} = \frac{\sqrt{E}}{2}$

۴. نتایج شبیه سازی

در این مقاله هدف کاهش احتمال خطاهایی است که زیست شناسان در حین توالی یابی DNA با آن برخورد دارند. این خطاها به صورت یک نویز سفید گوسی مدل شده است که در فرستنده به سیگنال ورودی اضافه می شود. با توجه به آنچه در فصول قبل گفته شد، ابتدا رشته ورودی DNA را به صورت صفر و یک مدل می کنیم. در قسمت شبیه سازی چون توالی خاصی از DNA مد نظر نیست برای حفظ کلیت مطلب، ورودی یک مجموعه از صفر و یک های تصادفی در نظر گرفته شده است. در این مقاله بدون از دست دادن کلیت، شناسایی باز آدنین را در توالی مد نظر داریم. به طوریکه به ازای هر ورودی 00، یک پیک در محل شماره باز آدنین خواهیم داشت. سپس مجموعه به دست آمده را با یک نویز گوسی سفید جمع شونده با توان نویز متغیر جمع می کنیم. درگیرنده همبستیگر را طراحی می کنیم. ضریب کاهش دامنه در همبستیگر به صورت am/m تعریف می شود که در آنm، شماره باز و am، متغیر تعریف شده است.

یکی از عوامل دیگری که در ایجاد خطا موثر است، روی هم افتادگی (over loading) پیک های گوسی است که این تداخل با شیفت مکان پیکها کاهش می یابد. مقدار شیفت با q_1 نشان داده شده است. با دقت در شکلهای۱1 تا ۱۵ بدست آمده در شبیه سازی متوجه می شویم که چند پارامتر در ایجاد خطا موثرند که با تغییر آنها می توان میزان خطای بوجود آمده در توالی یابی DNA را کاهش داد. مهمترین عامل در کم کردن خطا، کاهش توان نویز می باشد، به گونه ای که SNR افزایش یابد. این کار را می توان با بالا بردن دقت دستگاه و کاهش عوامل محیطی موثر در ایجاد این خطاها، انجام داد. با مقایسه شکل۱۱ با ۱۲و شکل ۱۳با ۱۴ که دارای q1 یکسان ولی am های متفاوت هستند متوجه می شویم که افزایش am سبب کاهش خطا می شود چرا که am /m ضریب میرایی محسوب می شود و افزایش am میزان این میرایی را بخصوص در پیکهای انتهایی کاهش می دهد و در نتیجه سطح زیر پیکهای حاصل از بازهای آخری مقدار بیشتری از λ خواهند داشت و ما پیکهای کمتری را از دست خواهیم داد. با مقایسه شکل۱۱با ۱۳ و شکل ۱۲ با ۱۴ که دارای am یکسان ولی q1 های متفاوت هستند به این نتیجه می رسیم که افزایش q1 یعنی شیفت پیکها در زمان به علت کاهش روی هم افتادگی پیکها سبب کاهش خطای توالی یابی می شود.

۵. نتیجه گیری و پیشنهادات

استفاده از تکنیکهای موجود در تئوری مخابرات برای استفاده در بیولوژی خصوصاً توالی یابی DNA روش جدیدی است که تاثیر زیادی در آشکار سازی و مدل کردن خطاهای حاصل از آن دارد. استفاده از این تکنیکها سبب افزایش سرعت و دقت در سیستمهای توالی یاب می شود.

برای افزایش دقت در سیستمهای توالی یاب می توان در بخش مدل کردن، بلاک دیاگرامی به نام حذف ISI داشت، که دراین صورت علاوه بر استفاده از فیلترهمسان، بلاکی با عنوان تخمین پیک و تخمین پالس مورد نیاز خواهد بود. در این مقاله تنها نویز سفید در نظر گرفته شده است، در صورتی که در سیستمهای واقعی نویز رنگی نیز بخش مهمی از نویزهای سیستم را تشکیل می دهد. برای بررسی دقیق خطاهای ایجاد شده در سیستم باید علاوه بر نویز سفید، نویز رنگی را نیز در کانال مخابراتی محسوب کرد.

تشکر

از اساتید گرانقدر جناب آقایان دکتر خادمی و دکتر ضمیری اعضای هیئت علمی گروه برق دانشگاه فردوسی مشهد که در تهیه این مقاله ما را یاری دادند کمال تشکر را داریم.

مراجع :

[1] Stephen William Davies “Application of Communication Theory to Automatic DNA Sequencing "PhD, Thesis, Dept. Elect. Computer. Eng. Univ. Toronto Canada, 1999

[2] Simon Haykin "Communication Systems" 1978

[3] http:// www.Electrophoresisanimated.htm

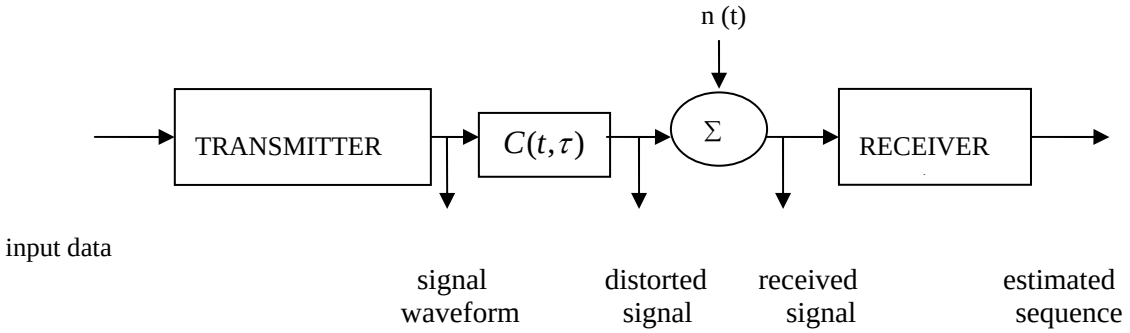
[4] http://core.biotech.hawaii.edu/g-dnaseq.htm

[5]محمود بهزاد "بیوتکنولوژی (مهندسی ژنتیک) " ۱۳۷۴

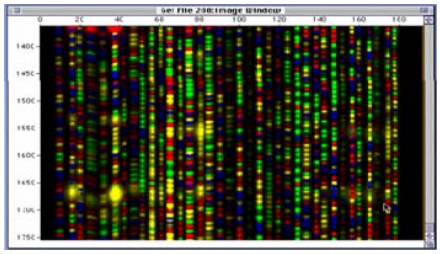
[6] Primerose,Sydny ترجمه طباطبایی یزدی ،محمد رضا نوری دلویی "بیوتکنولوژی مولکولی " مرکز ملی تحقیقات مهندسی ژنتیک و تکنولوژی زیستی ۱۳۷۲

[7] غلامرضا نورزاد "بیولوژی سلولی و بیولوژی مولکولی" ویرایش ۲ مشهد جهاد دانشگاهی .۱۳۷

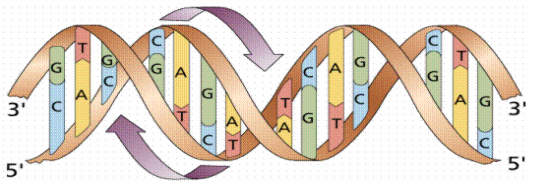
[8] جی ویلیامز –آ.اسکارلی-آ.والاس ترجمه : منصور مشرقی ،پرهام جبار زاده "مهندسی ژنتیک" ۱۳۸۱



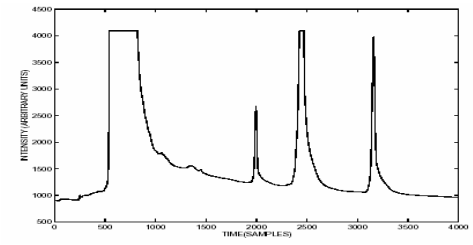
شکل ۱: بلوک دیاگرام سیستم مخابرات داده[1]



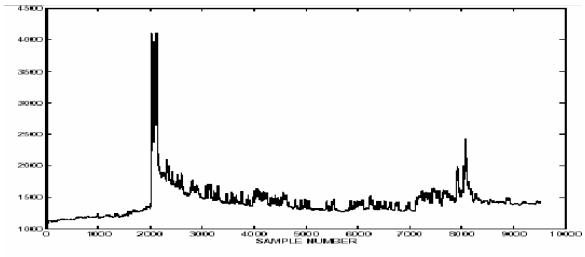
شکل ۳ آشکار سازی نوکلئوتیدهای فلورسنت شده با استفاده از لیزر[3].



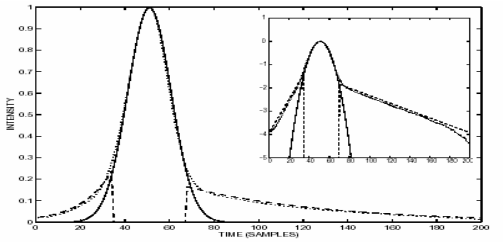
شکل ۲ : ساختار مارپیچDNA [4]



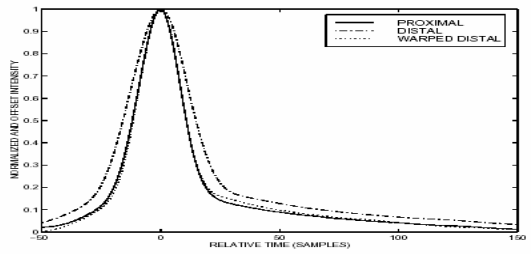
شکل۵ : پیکهای دور از مبدا و نزدیک مبدا قبل از انطباق [1]



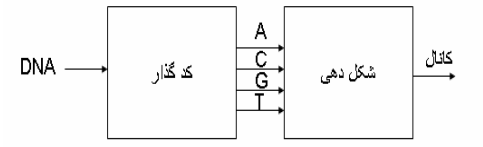
شکل۴: شگل حاصل از دستگاه توالی یاب[1]



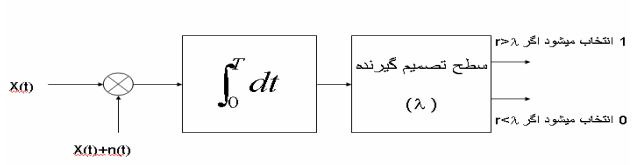
شکل ۶: تقریب نمایی از پیک دور از مبدا[1]



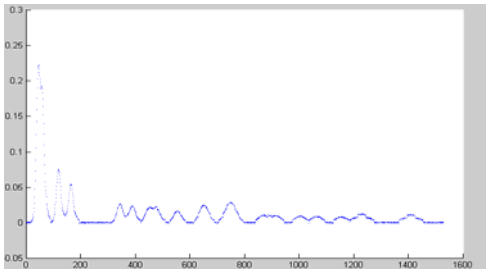
شکل ۷: پیک دور از مبدا (خط پر) و نزدیک مبدا (خط نقطه)[1]



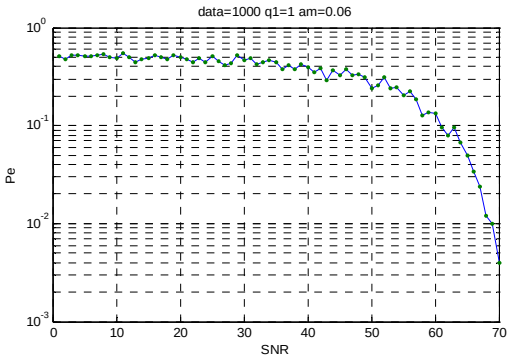
شکل ۸:کد گذاری بازها



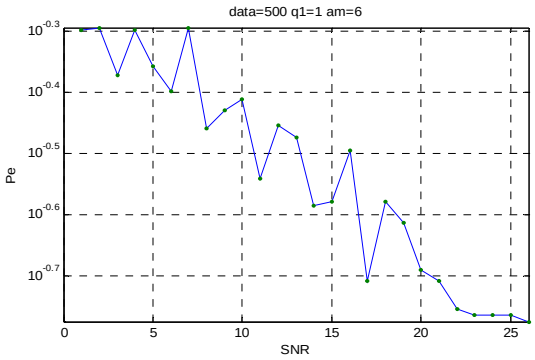
شکل ۹: مقایسه حاصل انتگرال با سطح آستانه



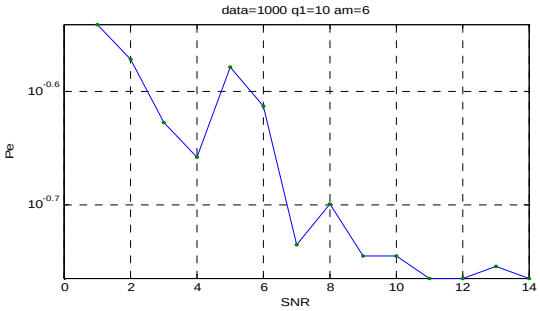
شکل۱۰ : مجموع ورودی و نویز با SNR= 70 dB



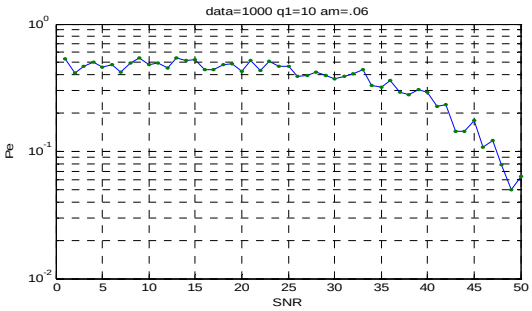
شکل ۱۱: تابع احتمال خطا برای ۵۰۰ با q1=1 ,am=0.06



شکل۱۲ : تابع احتمال خطا برای ۵۰۰ باز q1=1 , am=6



شکل۱۳ : تابع احتمال خطا برای ۵۰۰ باز q1=10 ,am=.06



شکل۱۴ : تابع احتمال خطا برای ۵۰۰باز q1=10 , am=6