روش تصمیم گیری پسخور در شناسایی زنجیره های DNA

سارا حاجی کاظمی ، حسین ضمیری و مرتضی خادمی

دانشگاه فردوسی مشهد، دانشکده مهندسی، گروه برق

E-mail: sarah.hajikazemi@gmail.com، hzamiri@ferdowsi.um.ac.ir ، mortezakhademi@yahoo.com

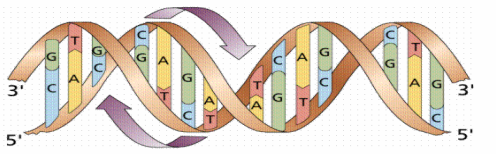
چکیده: زنجیره *DNA* (*deoxyribonucleic acid*) شامل اطلاعات ژنتیکی هر فرد می باشد، استخراج این اطلاعات به عنوان توالی یابی *DNA* (*DNA sequencing*) شناخته می شود. از آنجا که خطاهای کوچک در فرایند استخراج اطلاعات سبب ایجاد اشتباهات بزرگ در آزمایشات ژنتیکی می شود، دقت در طی فرایند توالی یابی برای کاهش خطاها بسیار حایز اهمیت می باشد برای مدلسازی این خطاها و کاهش احتمال خطا در طی این فرآیند از تئوری آشکار سازی در سیستم های مخابرات دیجیتال کمک گرفته می شود.در این مقاله روش آشکارسازی زنجیره های *DNA* با به کارگیری فیلتر منطبق تغییر پذیر با زمان و استفاده از روش تصمیم گیری پسخور ارائه شده است. شبیه سازی ها نشان می دهد که روش تصمیم گیری پسخور پیشنهادی دارای کارآیی خیلی بهتری در مقایسه با روش معمولی آشکارسازی زنجیره *DNA* است و احتمال خطا را به طور قابل توجهی کاهش می دهد.

کلید واژه ـ DNA ، توالی یابیDNA (DNA sequencing)، اکوالایزر تصمیم گیر با پسخور(Decision feedback equalizer)، سیستم های متغیر با زمان .

۱- مقدمه

فرآیند توالی یابی DNA در رأس پروژه ژنوم انسانی قرار دارد. ژنوم بیان کننده گروه کاملی از ساختار ژنتیکی (مجموعه ای از n کروموزوم) است که به وسیله یک سلول از یک ارگانیزم حمل می‌شود. دانشمندان امیدوارند که این پروژه انقلابی در علوم پزشکی و نحوه درمان بیماری ها ایجاد نماید. اهداف پروژه ژنوم انسانی عبارتند از تشخیص تقریبی ۲۰۰۰۰ تا ۲۵۰۰۰ ژن در ژنوم انسان، بیان توالی ۳ بیلیون جفت باز شیمیایی سازنده DNA انسان، ذخیره اطلاعات در Data base ها وبهبود روش آنالیز داده ها. توالی DNA ساختارهایی را برای هر آنچه که یک سلول انجام می دهد، از لحظه‌ی تولد تا مرگ شامل می‌شود. تغییرات در ساختار DNA به عنوان مثال می تواند شانس مبتلا شدن به بیماری را افزایش دهد و در قدرت مقابله بدن با بیماری و یا مقاومت بدن در برابر داروها تغییراتی ایجاد نماید. با توالی یابی دقیق DNA، دانشمندان قادر خواهند بود عامل بیماری ها را شناسایی و نحوه درمان آنها را بیابند.همان طور که در شکل ۱دیده می شود DNA به طور معمول از دو رشته (double strand) 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) وتیمین (T) می‌باشند. یک double strand از اتصال دو single strand که مکمل یکدیگرند تشکیل شده است. A و T مکمل و C و G مکمل یکدیگرند.

در بیولوژی از روش های مختلفی برای توالی یابی استفاده می شود که همه آنها شامل چند مرحله از جمله PCR (واکنش زنجیره ای پلیمراز) است.



شکل ۱: ساختار مارپیچ DNA

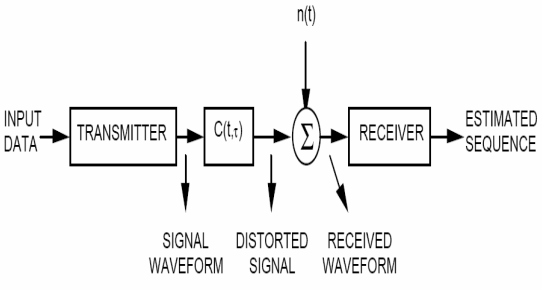
در فرآیند PCR، DNA به طور نمایی زیاد می شود.مرحله بعد توالی یابی به وسیله تئوری الکتروفورسیز توضیح داده می شود که در آن چاهکهای موجود بر روی ژل به عنوان غربال هایی برای مولکولها عمل می کند به طوری که حرکت انواع مختلف اسید نوکلئیک به طول آنها بستگی دارد. به این ترتیب بر اساس تفاوت در اندازه طول، DNA ها از هم جدا می شوند. از فلورسنت رنگی برای علامتگذاری نوکلئوتیدها استفاده می‌کنیم که همان طور که در شکل ۲مشاهده می شود با لیزر آشکارسازی می شوند خطاها در فرآیند توالی یابی DNA بر سه نوع می باشند.

خطای جابجایی[1] زمانی رخ می‌دهد که یک باز را به جای باز دیگر بخوانیم.خطای میان گذاری[2] زمانی رخ می‌دهد که در اثر اشتباه یک باز را اضافه بخوانیم. خطای حذف[3] زمانی رخ می‌دهد که در اثر اشتباه یک باز را نخوانیم. هر کدام از این خطاها می‌توانند بر اثر دقت پایین دستگاه توالی یاب رخ دهند. آشکارسازی توسط دستگاه توالی یابی دارای همه خطاهای فوق است. مخابرات داده بر پایه انتقال و دریافت رشته اطلاعات با احتمال کمی خطا استوار است. در شکل ۳ عناصر یک سیستم مخابرات داده نشان داده شده است. [۱]

در ابتدا، داده های ورودی وارد فرستنده شده و یک سیگنال آنالوگ در ورودی کانال حاصل می‌شود. عبور از کانال دو اثر مهم دارد. ابتدا شکل موج توسط پاسخ ضربه کانال خراب می‌شود.در مرحله بعد نویزبه اطلاعات اضافه می‌شود. گیرنده، خطاهای آشکارسازی را با تخمین زدن، حذف ISI و میانگین گیری در زمان، برای کاهش تأثیر نویز، کاهش می‌دهد. در این مقاله، توالی یابی DNA به صورت وابسته به نویز و رویهم افتادگی سیگنال که یک توالی اطلاعاتی را بیان می کند، مدلسازی شده است. مخابرات داده نیز برای استخراج یک توالی اطلاعاتی از حالت نویزی و روی هم افتادگی سیگنال ها به کار رفته است. به طورقطع شباهت قوی بین مخابرات داده و توالی یابی DNA وجود دارد. در بخش های بعدی ابتدا یک مدل از سری زمانی DNA و سپس سیستم مخابراتی استفاده شده ارائه می شود و نهایتاً نتایج شبیه سازی مورد بررسی قرار خواهد گرفت.

[1] Substitution
[2] Insertion
[3] Deletion

شکل ۲: آشکار سازی نوکلئوتیدهای فلورسنت شده با لیزر

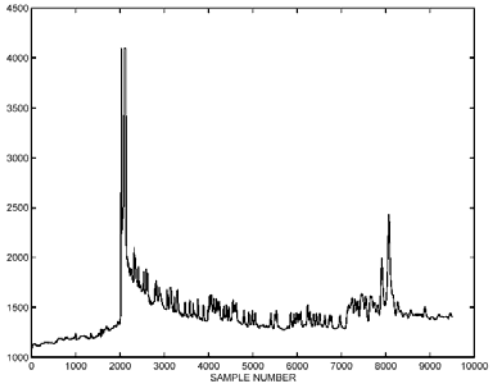


شکل ۳: شکل ۱، بلوک دیاگرام سیستم مخابرات داده

۲- سری زمانی DNA

در این بخش یک مدل آماری از سری زمانی DNA به دست آمده از دستگاه توالی یاب معرفی می شود. شکل پیک و پارامترها مدل شده است و نویز کل سیستم یک نویز گوسی جمع شونده فرض می‌شود.

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 می‌شود. با قدرت تفکیک بیشتر برای چهار بازمتفاوت، مدل سری زمانی به صورت رابطه (۱) پیشنهاد می‌شود. [۱]



شکل۴: شکل حاصل از دستگاه توالی یاب [۱]

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

که n نوع باز را نشان مـی دهـد. انـدیس k شـماره نمونـه و جمع روی موقعیت توالی بـاز i ام و N_b تعـداد کـل بازهـایی است که توالی یابی مـی شـوند. $\delta_{n,xi}$ در تـشخیص ایـن کـه آیا x_i ,یعنی باز در موقعیت i ام توالی، همـان نـوع کانـال n است، استفاده می شود. به این ترتیب که، در صورتی کـه n و χ_i برابر باشـند، یـک ودر غیـر ایـن صـورت صـفر اسـت. متغیرهـای t_i و a_i لـرزش زمـانی و نوسـان دامنـه را مـدل می‌کنند و درنهایت یک نویز جمع شـونده $\{n_k\}$ داریـم کـه نوسان سطح زمینه را معرفی می کنـد. الکتروفورسـیز یـک مولکول خالص وقتی که مولکول هایDNA به اندازه کـافی دوراز زمان بارگذاری حرکت کند، باید منجر بـه یـک پیـک گوسی شود. اگرچه در نتایج مشاهده شده از دستگاه تـوالی یاب، شکل پیک بسیار پیچیده تر از گوسی است. در شکل ۵ چهار نوع پیک نمونه نشان داده شده است. [۱]

در شکل ۶ خط پیک هر ناحیه حرکت کرده تـا پیـک دور از مبداء با پیک نزدیک مبدا تطبیق یابد.

تنها تفاوت بارز در پیک ها، بسطِ افتن پیک های انتهایی در زمان است. پیک ها می توانند با فرمول(۲) مدل شوند.

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

k اندیس نمونه، t_i زمان پیک، g_1 شـکل پـالس (پیوسـته) وقتـی عـرض پیـک واحـد باشـد و $P_\omega(t_i)$ نـشان دهنـده وابستگی عرض پیک به زمان پیک است که به صـورت زیـر مدل می شود[۱] :

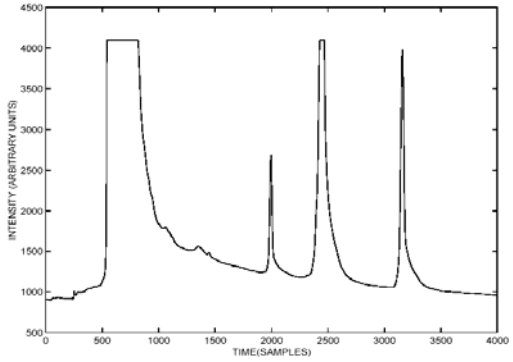
$$p_w(t_i) = 15.08 + 0.0326(t_i - 1) \qquad (۳)$$

در شـکل ۷، شـکل پـالس بـه سـه ناحیـه مجـزا تقـسیم می شودکه لوب اصلی گوسی و دو دامنه نمایی است. [۱]

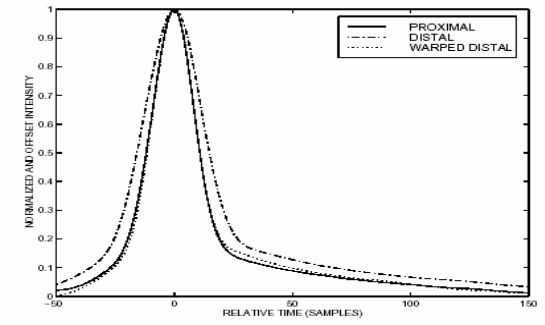
شکل پالس کلی با عرض واحـد بـرای مجموعـه داده تـوالی یابی، فرمول (۴) را نتیجه می‌دهد،

$$g_1(l) = \begin{cases} 7.89e^{4.8l}, & l < -0.86 \\ e^{-(1.67l)^2}, & -0.86 < l < 0.64 \\ 6.89e^{-1.96l}, & 0.64 < l \end{cases} \qquad (۴)$$

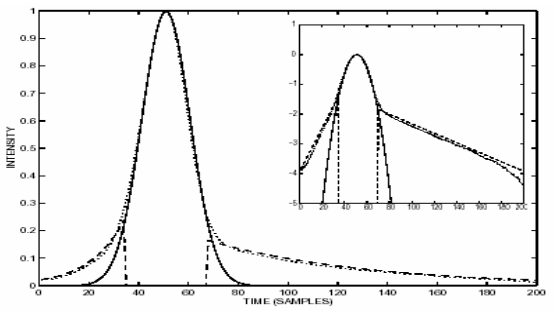
برای مـدل کـردن و شـبیه سـازی سـاختار DNA از یـک سیستم مخابراتی اسـتفاده مـی کنـیم کـه در بخـش ۳ بـه توضیح آن می‌پردازیم.



شکل ۵: پیکهای دور از مبدا و نزدیک مبداء قبل از انطباق



شکل ۶: پیک دور از مبدا (خط پر) و نزدیک مبداء (خط نقطه)



شکل ۷: تقریب نمایی از پیک دور از مبداء

۳- مدل سیستم مخابراتی

همان طور که در شکل ۸ نشان داده شده است، در فرستنده به هر باز یک کد اختصاص می دهیم و بازهـای DNA را بـا این کدها شناسایی می کنیم. A با 00 ، C با 01 ، G بـا 10 و T با 11 نشان داده شده است.

در محل هر باز پیکی رخ می دهد که تنها تفاوت پیک ها در عرض پالس و دامنه است. خطاهای ناشی از میان گـذاری و حذف ناشی از نویز جمع شونده ای است کـه آن را سـفید و گوسی مدل می‌کنند. همان طور که در شکل ۹ نـشان داده

شده است، مطابق رابطه (۳) با افزایش تعداد نمونه ها، پالسها در حوزه زمان گسترده شده و دامنه آنها نیز کاهش می‌یابد. در این شکل تداخل سیگنال ها و تأثیر هر سیگنال بر روی نمونه بعدی به خوبی مشهود است.

بدیهی است که پس از جمع پالس ها در هر زمان، در هر نقطه سطح انرژی افزایش می یابد. در شکل ۱۰ دیده میشود که در نمونه های انتهایی رشته ISI ایجاد شده به همراه نویزی که در کانال به پالس ها افزوده می‌شود، آشکار سازی را بسیار مشکل می سازد. با توجه به دوره هر پالس در این جا می‌توان تخمین زد که هر پالس با دو نمونه پیش از خود تداخل می کند. برای رفع این مشکل و کاهش احتمال خطا از اکوالایزر تصمیم گیرنده با پسخور استفاده می کنیم.

شکل ۱۱ بلوک دیاگرام یک سیستم آشکارساز به روش پسخور را نشان می‌دهد. سیگنال r(t) سیگنال دریافتی است که جهت آشکارسازی زنجیره DNA به کار می رود. شکل ۱۰ یک نمونه نوعی از سیگنال وروی r(t) را نشان میدهد که ناشی از وجود بازهای نشان داده شده در شکل ۹ می باشد.

همان طور که شکل ۱۱ نشان می دهد پهنای هر پالس که ناشی از وجود یک باز است، می تواند در بازه مجاور گسترش یابد. بنابراین ISI ناشی از دو باز در دو طرف باز باعث خطا در آشکارسازی می شود. نظر به این که تصمیم گیر با پسخور فقط توانایی حذف ISI ناشی از گذشته را دارد بنابراین $H(l,k)$ بصورت زیر مدل سازی می‌گردد:

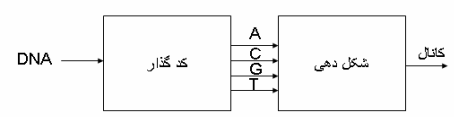
$$H(l,k) = h_{1,k}\,q^{-1} + h_{2,k}\,q^{-2} \qquad (۵)$$

همان طورکه در رابطه (۵) دیده می شود ضرایب فیلتر پسخور وابسته به آشکارسازی باز k ام است که این ضرایب از رابطه زیر به دست می‌آیند.

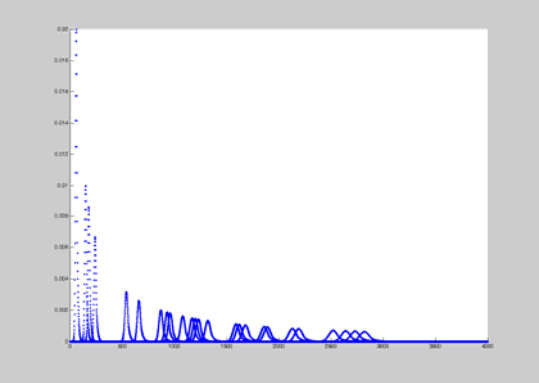
$$h_{1,k} = \int_{-\infty}^{+\infty} g(t,k)g(t,k-1)dt \qquad (۶)$$

$$h_{2,k} = \int_{-\infty}^{+\infty} g(t,k)g(t,k-2)dt$$

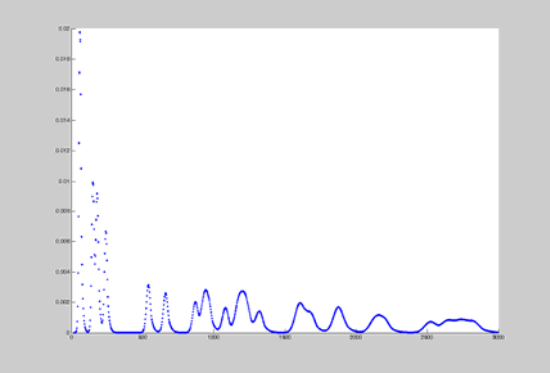
از طرفی نظر به این که انرژی $g_{t,k}$ برایk های گوناگون متفاوت است، $\lambda(k)$، سطح آستانه تصمیم گیری نیز متغیر با زمان است. هر گاه $\lambda(k)$ سطح آستانه تصمیم گیری باشد احتمال خطا برابر است با:



شکل ۸: کدگذاری بازها



شکل ۹ : تأثیر کانال بر روی پالس های ورودی برای ۵۰باز



شکل ۱۰: مجموع ورودی و نویز برای ۱۰۰ باز با SNR=30dB

$$P_e = \frac{P_1}{2}\,erfc\left(\frac{\sqrt{E_k} - \lambda(k)}{\sqrt{N_0}}\right) + \frac{P_0}{2}\,erfc\left(\frac{\lambda(k)}{\sqrt{N_0}}\right) \qquad (۷)$$

که در آن E تغییر پذیر با زمان بوده و آن را به صورت E_K در نظر می گیریم که از رابطه زیر حاصل می‌شود:

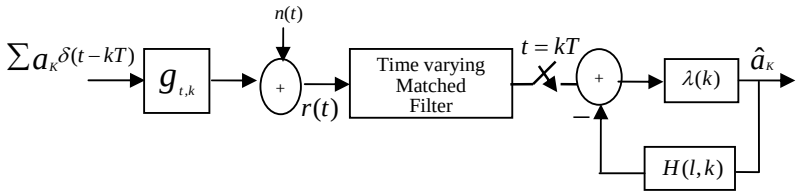
$$E_K = \int_{-\infty}^{+\infty} g_{t,k}^2\,dt \qquad (۸)$$

چنانچه p_1 احتمال وجود باز و p_0 احتمال عدم وجود باز باشد برای مینیمم سازی خطا می توان نشان داد که:

$$\lambda_{Opt}(k) = \frac{N_0}{2\sqrt{E_K}}\left(Ln\left(\frac{p_0}{p_1}\right) + \frac{E_K}{N_0}\right) \qquad (۹)$$

هر گاه $p_0 = p_1$ باشد، داریم:

$$\lambda_{opt}(k) = \frac{\sqrt{E_k}}{2} \qquad (۱۰)$$



شکل ۱۱: بلوک دیاگرام سیستم آشکارساز با تصمیم گیری پسخور

۴- نتایج شبیه سازی

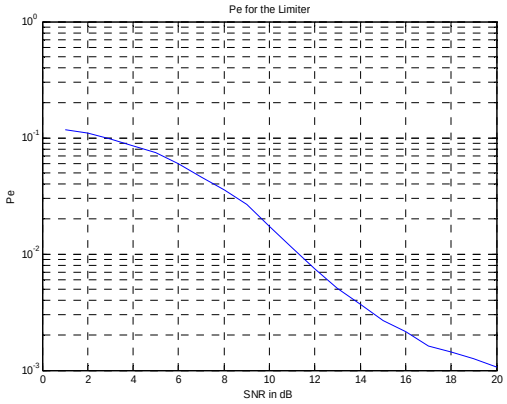
در شبیه سازی ها ابتدا رشته ورودی DNA را به صورت 0 و 1 مدل می کنیم. چون توالی خاصی از DNA مد نظر نیست برای حفظ کلیت ، ورودی را یک مجموعه از 0 و 1 های تصادفی که دارای توزیع یکنواخت می باشد در نظر گرفته می‌گیریم.

در این جا بدون از دست دادن کلیت، شناسایی باز آدنین را در توالی مدنظر داریم. به طوری که به ازای هرورودی 00 یک پیک در محل شماره باز آدنین خواهیم داشت.

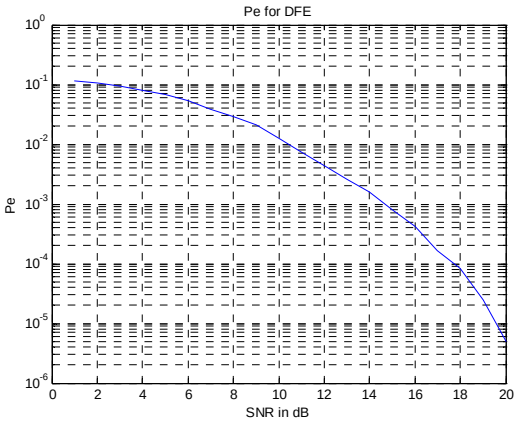
شکل های ۱۲ و ۱۳ به ترتیب احتمال خطای آشکارسازی زنجیره DNA برای آشکارسازی معمولی (بدون استفاده از DFE) و با استفاده از روش تصمیم گیری پسخور را نشان می دهد. همان طور که ملاحظه می گردد احتمال خطای آشکارسازی پیشنهادی با پسخور به طور قابل توجهی کمتراز روش معمولی آشکارسازی زنجیره DNA می باشد.

۵- نتیجه گیری

در این مقاله با استفاده از تئوری آشکارسازی در سیستم های مخابرات داده جهت آشکارسازی زنجیره DNA برپایه گیری با پسخور ارائه گشت. نظر به این که دامنه پالس های ناشی از وجود بازها در زنجیره DNA با افزایش طول DNA کاهش می یابد. آشکارسازی پیشنهادی با استفاده از یک فیلتر منطبق تغییر پذیر با زمان و نیز یک سیستم تصمیم گیر که سطح آستانه آن نیز تغییر پذیر با زمان است و برای مینیمم سازی احتمال خطا طراحی گشته است شده است، صورت می پذیرد.فیلتر FIR تغییر پذیر با زمان که در حلقه تصمیم گیری پسخور قرار دارد، باعث کاهش ISI میگردد.شبیه سازی ها نشان داد که احتمال خطای آشکارسازی با روش پیشنهادی با تصمیم گیری پسخور به طور قابل توجهی نسبت به روش معمولی آشکارسازی کاهش می‌یابد.



شکل ۱۲: تابع احتمال خطا(probability of error,Pe) برای ۱۰۰ باز بدون استفاده ازروش آشکارسازی پسخور



شکل ۱۳: تابع احتمال خطا (probability of error,Pe) برای ۱۰۰ باز با استفاده ازروش پسخور پیشنهادی

۶- مراجع

[1] Simon Haykin, Communication Systems, 1978.
[2]Stephen William Davis, Application of Communication Theory to Automatic DNA Sequencing, PhD thesis, 1999.
[3]Rodney Andrew Kennedy, Operational Aspects of Decision feedback equalizers, B.E. (Hons.) UNSW,1988.

[4]Claes Tidestav, The Multivariable Decision Feedback Equalizer Multiuser Detection and Interference Rejection, Uppsala University Signals and Systems Dept,1999.
[5]http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[6]http://www.blc.arizona.edu/Molecular_Graphics/DNA_Structure/DNA_Tutorial.HTML#Components
[7]http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/pg1.html
[8]http://www.bergen.org/AAST/Projects/Gel/intro.htm
[9]http://content.febsjournal.org/cgi/content/full/269/15/3632#F
[10]http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/gels/principles.html