

Chemical Synthesis of Oligonucleotides

Mark A. Behlke, M.D., Ph.D. and Eric J. Devor, Ph.D.
Molecular Genetics and Bioinformatics
Integrated DNA Technologies

Introduction

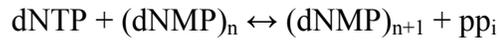
It can fairly be said that synthetic oligonucleotides are the fuel that drives the engine of molecular biology. Nearly every technique in use today in molecular biology employs chemically synthesized DNAs or RNAs. This includes PCR, Real-Time PCR, DNA sequencing, site directed mutagenesis, single-nucleotide polymorphism (SNP) assays, microarrays, and the rapidly expanding world of small RNAs. Unlike other reagents, however, oligonucleotides are not stock items that can be stored, warehouse fashion, until needed. Each oligonucleotide is custom made according to the specific needs of the individual researcher and purified according to the application for which it is intended.

In recent years, advances in oligonucleotide synthesis chemistries as well as in purification and quality control technologies have led to substantial increases in both quality and yield and to substantial decreases in cost. At the same time, developments in oligonucleotide applications have spurred a dramatic expansion of the range of available modifications. As technological advances in applications have ushered in an era of high throughput analyses, so, too, advances in oligonucleotide synthesis, purification, and QC have led to the establishment of high throughput manufacturing of the full range of possible syntheses such as those available at Integrated DNA Technologies.

Here, we will present the fundamentals of chemical synthesis of oligonucleotides as it is practiced today and will present synopses of some of the more basic issues arising from such syntheses including those bearing on scale, yield, and some of the available modifications. We will also provide links to additional readings relevant to other aspects of oligonucleotide synthesis including fluorescence, nuclease resistance, chimeric constructs, purification, and QC. We begin with a brief history of *in vitro* synthesis as a way of setting the context of the high throughput synthesis environment in which we now operate.

A History of Chemical Synthesis

Elucidation of the Genetic Code, accomplished in full by 1966, is an example of brute force applied to a delicate problem. One of the main figures in that story was Professor H. Gobind Khorana, a biochemist at the University of Wisconsin. It was understood that, in nature, the formation of the phosphate linkages in DNA was catalyzed enzymatically in a reversible reaction:



That is, a deoxynucleotide triphosphate was added to a growing deoxynucleotide monophosphate polymer by a polymerase in the presence of Mg^{++} and that the forward reaction resulted in the $n+1$ polymer and a pyrophosphate. Once this was understood, several investigators, notably Khorana, began to try ways to accomplish this synthetically. Fortunately, many of these investigators were already well versed in the chemistry involved in the synthesis of polypeptides, a technique that had been around for many years by the late 1950's. This very successful chemistry, shown graphically in figure 1, utilized a foundation of a single amino acid bound to a **solid support**, in this case a resin bead, and a **protected** amino acid in the presence of a condensing agent. The

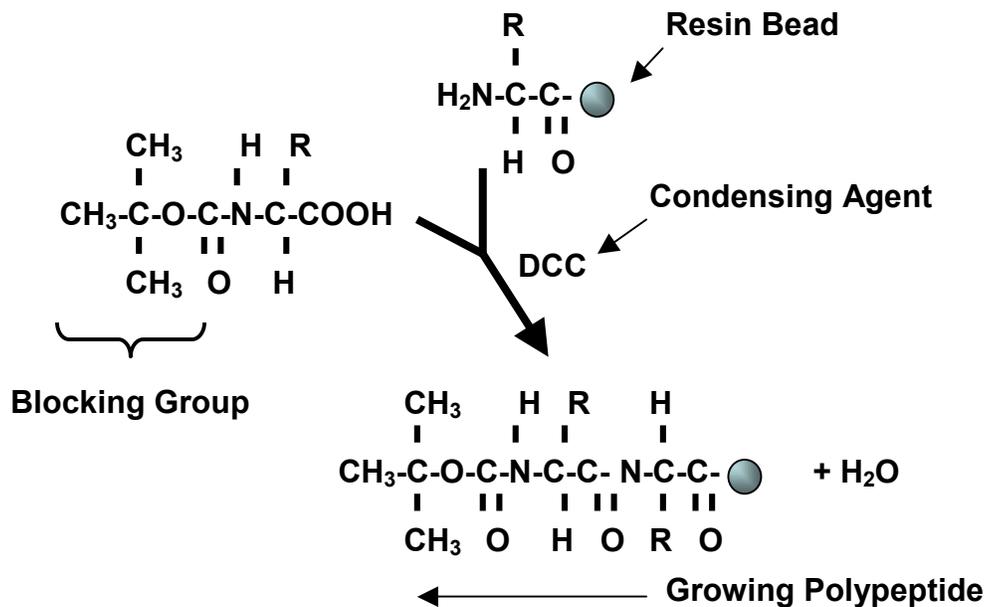


Figure 1. The basic scheme of polypeptide synthesis from a solid support (resin bead). The final deprotection step removed the blocking group from the coupled amino acid and permitted the next coupling reaction to proceed in the presence of DCC.

condensing agent, dicyclohexylcarbodiimide (DCC), catalyzed the peptide linkage. The reaction would link the protected amino acid to the solid support in only one way and the resulting di-peptide was then **deprotected** for the addition of the next amino acid. In this way, the synthetic polypeptide grew at the direction of the synthesizing chemist one link at a time through a repeated cycle of steps.

Khorana and his group at the Institute for Enzyme Research at the University of Wisconsin began experimenting with DCC and various solid supports and protecting groups on both ribonucleotides and deoxyribonucleotides. The scheme they developed involved three different protecting groups surrounding the nucleotide (figure 2). The integrity of the ring structure of the base was protected by a **benzoyl** or an **isobutyryl** group, the oxygen on the 3' carbon of the sugar was prevented from forming a reactive

hydroxyl until needed by the addition of an **acetyl** group, and the 5' carbon was blocked by a **trityl** group. In the synthesis of the trityl group Khorana chose to use a soluble polymeric support. In the case of polypeptide synthesis the support was a resin bead but Khorana settled on a polystyrene bead for nucleic acid synthesis. Synthesis of the trityl would begin with attaching a benzene ring to the polystyrene bead. Next, a second ring was attached via a Friedel-Crafts reaction which is the acetylation (or alkylation) of aromatic compounds by aluminum chloride. This structure was then exposed to a Grignard reagent, an extremely reactive class of compounds used in the synthesis of hydrocarbons. Finally, the nearly completed trityl group was treated with acetyl chloride resulting in a polystyrene methoxytritylchloride species that could be easily attached to the 5' carbon of the pentose sugar (figure 2).

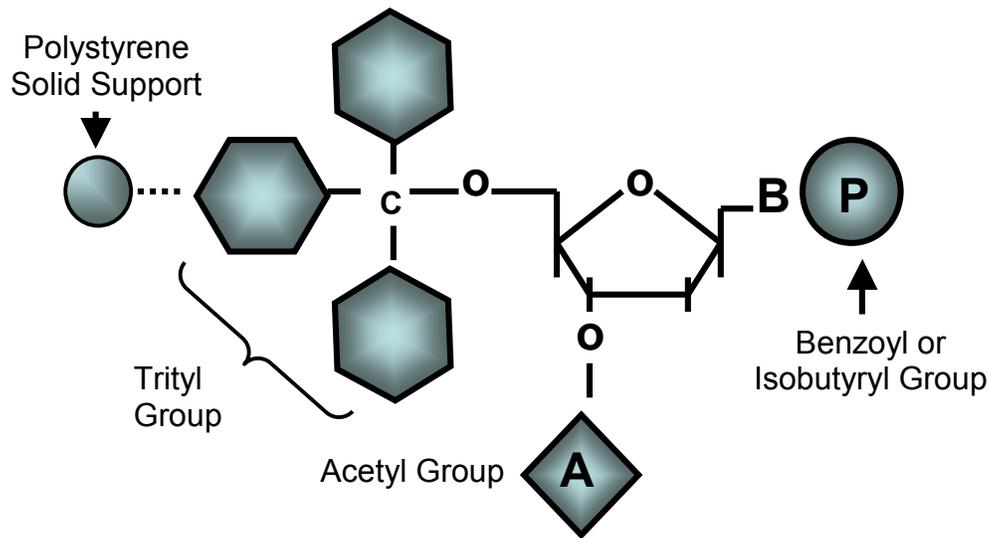


Figure 2. Starting point structure of an oligonucleotide synthesis by the Khorana group. Subsequent couplings would proceed through the 3' carbon after removal of the acetyl group. All nucleotides added from this point would have a reactive hydroxyl in place of the trityl group but would retain the other two protecting groups.

By placing the solid support on the trityl group and using the DCC condensing reagent, the synthesis of an oligodeoxynucleotide proceeded in the same 5'→3' direction as in nature. The nucleotide to which the trityl was attached became the anchor of the synthesis. The oxygen on the 3' carbon of the sugar was deprotected leaving the reactive hydroxyl group reading for **coupling** and protected nucleotide without a trityl was added via the condensing reaction of DCC. This method was tedious and only modestly efficient but it led to two major breakthroughs. First, Khorana and his group were able to synthesize oligoribonucleotides that were used to confirm the Genetic Code. Second, the Khorana-led team at Wisconsin announced in 1967 their intention to use this chemistry to synthesize a gene. Their attempt to do this was “catalyzed” by the discovery in that year of the enzyme DNA ligase. Khorana realized that they could synthesize overlapping, complementary oligodeoxynucleotides and assemble the gene using the ligase. In 1970 they published the first completely synthetic gene, the 77bp yeast tRNA_{Ala} gene (Agarwal et al., 1970).

One final note about the accomplishments of the Khorana group, which included Khorana receiving the 1968 Nobel Prize in Physiology or Medicine. In their 1970 synthetic gene paper in *Nature*, they note, “Unpublished experiments by two of us have given encouraging results on the use of **DNA polymerase** for replication of the gene in the presence of suitable **primers**.” (Agarwal et al., 1970: 30, emphasis mine). There has been a great deal of both myth and litigation surrounding the invention of the polymerase chain reaction. It is difficult to believe that this sentence went unnoticed from 1970 to the mid-1980’s when PCR burst upon the scene. It would be like suggesting that the famous comment from Watson and Crick that, “It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.” (Watson and Crick, 1953: 738) was ignored by all but the authors.

Contemporary Oligonucleotide Synthesis

In spite of the successes of researchers like Khorana, chemical synthesis of oligonucleotides remained labor intensive and inefficient throughout the 1970’s and into the early 1980’s. In 1983 a breakthrough was achieved in synthesis chemistry that made it possible to make longer and longer oligonucleotides and to make them much more efficiently. The new synthesis process was based upon the use of **phosphoramidite monomers** and the use of **tetrazole** catalysis (McBride and Caruthers, 1983). A phosphoramidite monomer is a very different synthesis unit compared to its predecessors. First of all, the trityl group is not used to link to the solid synthesis support. Rather, it is a removable protecting group. The link to the solid support is made through the 3’ carbon and synthesis proceeds 3’ to 5’ rather than the 5’ to 3’ synthesis used previously (Figure 3A). The solid support used is a 5 micron controlled pore glass bead (CPG). This bead has a surface with holes and channels and it is in these that the protected nucleotide is attached (figure 3B).

Phosphoramidite synthesis begins with the 3’-most nucleotide and proceeds through a series of cycles composed of four steps that are repeated until the 5’-most nucleotide is attached. These steps are **deprotection, coupling, capping, and stabilization**. In the classic deprotection step the trityl group attached to the 5’ carbon of the pentose sugar of the recipient nucleotide is removed by trichloroacetic acid (TCA) leaving a reactive hydroxyl group. At this stage the next phosphoramidite monomer is added. It is here that the advent of **tetrazole** activation replaces the use of condensing agents like DCC. Berner et al. (1989) showed that tetrazole, a weak acid, attacks the coupling phosphoramidite nucleoside forming a tetrazolyl phosphoramidite intermediate. This structure then reacts with the hydroxyl group of the recipient and the 5’ to 3’ linkage is formed (figure 4). The tetrazole is reconstituted and the process continues. The use of tetrazole increased coupling efficiency to greater than 99% and, with this, opened the way for longer and longer oligonucleotides to be synthesized.

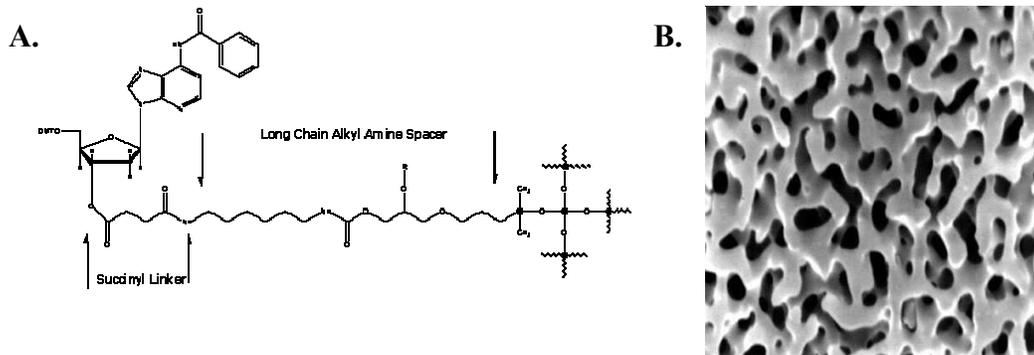


Fig. 3. A. A protected nucleoside attached to a CPG. B. An electron photomicrograph of the surface of a CPG bead. The scale of this picture is 10 millionths of an inch square.

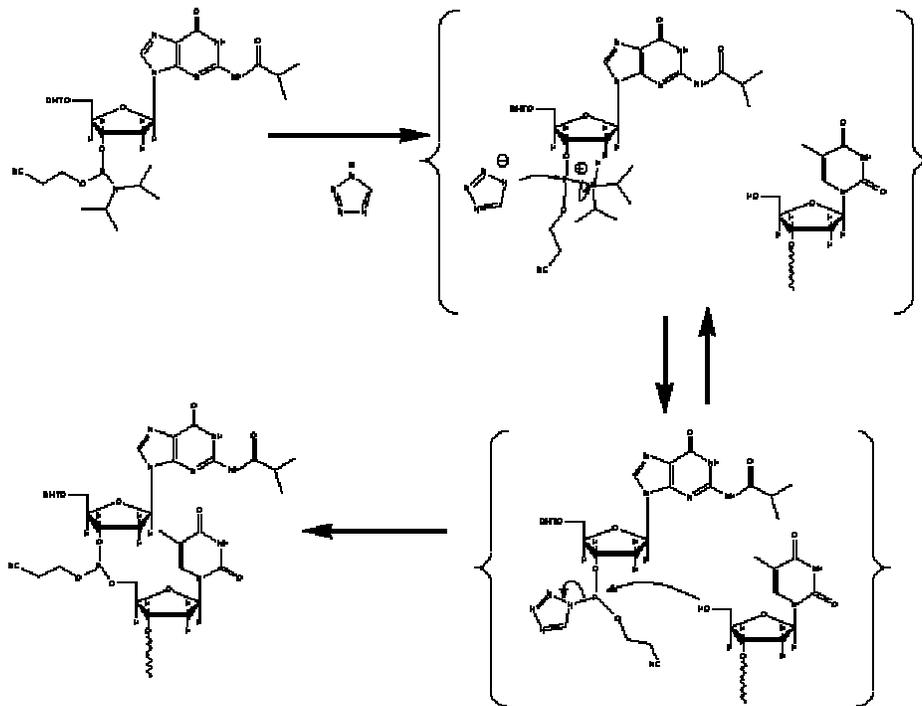


Figure 4. The pathway of tetrazole phosphoramidite-intermediate coupling. The phosphoramidite is introduced in the presence of tetrazole which protonates diisopropylamine leading to the formation of the tetrazole phosphoramidite intermediate. Coupling of the intermediate to the growing oligonucleotide is the final step which returns the tetrazole to its original state.

While the increased efficiency afforded by the advent of tetrazole phosphoramidite-intermediate coupling was a major advance in oligonucleotide synthesis, it was still a chemical process and there was a finite failure rate. A coupling failure results in an oligonucleotide still having a reactive hydroxyl group on its 5'-most end. If this were to remain freely reactive, it would be able to couple in the next round and the result would be a missing base in the synthesis. Thus, coupling failures had to be removed from further participation in the synthesis. This is classically accomplished by adding an

acetylating reagent composed of acetic anhydride and N-methyl imidazole. This reagent reacts only with free hydroxyl groups to irreversibly **cap** the oligonucleotides in which coupling failed.

Once the **capping** step is accomplished the last step in the cycle is **oxidation** which stabilizes the phosphate linkage between the growing oligonucleotide chain and the most recently added base. Again, in the classic phosphoramidite synthesis, this step is carried out in the presence of Iodine as a mild oxidant in tetrahydrofuran (THF) and water. The Water acts as the oxygen donor and the iodine forms an adduct with the phosphorus linkage. The adduct is decomposed by the water leaving the phospho-triester bond stabilized.

Following this final step the cycle is repeated for each nucleotide in the sequence. At the end of the synthesis the oligonucleotide exists as, say, a 25-mer with the 3' end still attached to the CPG and the 5' end protected with a trityl group. In addition, there are protecting groups on three of the four bases. These are present to maintain the integrity of the ring structures of the bases. The protecting groups are benzoyl on A and C and N-2-isobutyryl on G (figure 5). Thymidine needs no protecting group. The completed synthesis is **cleaved** off the CPG and then **deprotected** leaving a hydroxyl on both the 3' and 5' ends. At this point the oligo is **deprotected** and exists as a functional single-stranded DNA molecule. However, deprotection removes the protecting groups but they remain with the oligonucleotide as organic salts that must be removed. The process of removing these contaminants is called **desalting**.

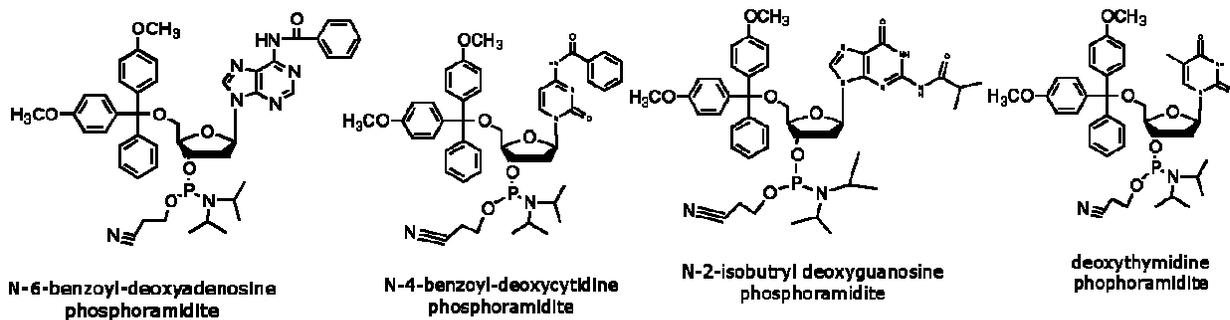


Fig. 5. Structures of the four nucleoside phosphoramidite monomers. The benzoyl and isobutyryl protecting groups on the A, C, and G monomers are shown.

Synthesis Scale and Yield

The difference between **synthesis scale** and **synthesis yield** can never be overstated. Synthesis scale refers to the amount of starting material that is composed solely of the 3' most nucleotide chemically linked to the CPG. Synthesis yield refers to the cumulative loss of mass due to coupling failures and subsequent purification, if any. This cumulative loss of mass is unavoidable because we are dealing with chemistry and physics not magic. Oligonucleotides must be assembled and they must, in many cases, be purified. These processes come with a cost that, in the contemporary synthesis world, is minimal and far outweighed by the benefits.

We noted above the phenomenon of **coupling failures** and the subsequent chemical process of **capping**. In contemporary phosphoramidite oligonucleotide synthesis coupling failure rates should be 0.5% or less per coupling event. This means that there is a greater than 99.5% success rate at each coupling step. However, this results in a cumulative population of failures such that, for a 25-mer, the percent full-length product can be estimated as $(0.995)^{24} = 0.8867$, or 88.67%. The exponent is (n-1) because the very first base on the 3' end is given as a result of the CPG. Thus, while there are 25 nucleotides there are 24 couplings. Similarly, for a 35-mer, the function is $(0.995)^{34} = 0.8433$ and for a 50-mer it is $(0.995)^{49} = 0.7822$. Additional information about the nature of coupling failures can be found in Temsamani et al. (1995) and Hecker and Rill (1998). For PCR, most primers are in the 20- to 28-mer range. This means that a standard synthesis will be between 85% and 90% full-length. PCR is a very forgiving process in that the overwhelming mass of full-length product will be sufficient to produce the amplicon of interest in an equally overwhelming mass. Thus, for PCR, there is little need for purification beyond that of desalting that removes the organic salts generated during the final deprotection process.

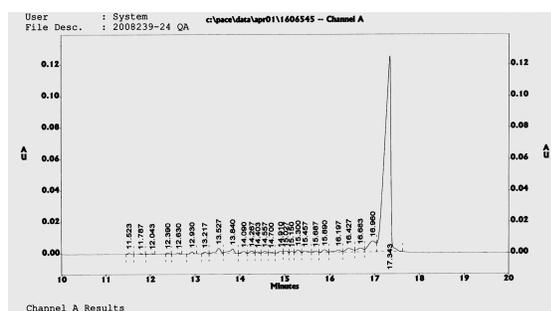
Subsequent **purification** of an oligonucleotide synthesis is application-dependent. Longer syntheses are usually not used for PCR. Longer syntheses are mainly used for cloning and hybridization applications. There, the non-full-length mass is important as it might interfere with those applications. Similarly, modified oligonucleotides; i.e., those syntheses that involve non-standard bases, fluorescent dyes, linkers, etc. should be subjected to additional purifications. There are two methods for subsequent purification. The first is **polyacrylamide gel electrophoresis** (PAGE) and the second is **high performance liquid chromatography** (HPLC). The choice of which purification method is used is largely dictated by the oligonucleotide itself. If the oligonucleotide is unmodified and the only issue is removal of the capped products, PAGE remains the most efficient means of purification. High percentage acrylamide gels will permit isolation of full-length product from all shorter species with great efficiency. The full-length product is simply excised from the gel and eluted from the gel slice. There is one drawback. As noted, coupling failures result in the loss of mass relative to the initial **synthesis scale**. This is unavoidable. PAGE purification also results in an unavoidable loss of mass because it is physically impossible to recover every bit of full-length product from a gel slice. Thus, a PAGE-purified 50-mer will present with substantially lower mass relative to the starting synthesis scale than will a desalted 25-mer. However, given that the applications for which a 50-mer is utilized are likely to be much more sensitive to non-full-length species, the loss of mass exchange for high purity is, or, at least, should be, an easy choice.

The other purification method commonly used, HPLC, comes in two basic forms; **reverse-phase** and **ion-exchange**. The choice of HPLC, and, specifically, which type of HPLC, is determined by the synthesis. HPLC purification is usually reserved for oligonucleotides that have been modified in some way. Again, this can be the addition of a linker or spacer, a non-standard base or bases, or fluorescent molecules. A 25-mer oligonucleotide will have a specific charge and a specific affinity, or lack of affinity, for

particular solvents. If the modified oligonucleotide can be better separated from the unmodified oligonucleotide on the basis of the change in net charge induced by the modification, then **ion-exchange HPLC** will be the method of choice. If the modified oligonucleotide can be better separated from the unmodified oligonucleotide on the basis of the change in solvent affinity, then **reverse-phase HPLC** will be the method of choice. Here, again, there will be an unavoidable loss of mass due to purification but, again, this should be more than offset by the gain in purity.

Synthesis and purification is not the end of the process, however. Synthesizing and purifying an oligonucleotide does not provide information about how good the oligonucleotide is. Of course, the ultimate court is the reaction tube but there are two methods by which the oligonucleotide can be tested for quality prior to its use in an experiment. These methods are **capillary electrophoresis (CE)** and **mass spectrometry**, here referred to as MALDI (see Additional Readings on CE and Mass Spectrometry for more information on these techniques). Both of these methods provide excellent first order data about synthesis quality. Capillary electrophoresis requires a small amount of the final synthesis product that is then subjected to a constant electrical field in a hair-thin capillary. As the product migrates in the capillary it is separated into component sizes in a manner exactly like gel electrophoresis. The fragments will migrate past an optical window and an ultra violet beam detector will assess the density of the fragments. The “trace” that is produced is composed of a series of peaks corresponding to material densities flowing past the detector (figure 6). This density profile is then made quantitative by establishing a base line and integrating the area under the individual peaks. The purity of the product is, then, the ratio of the main peak to the total area under all peaks. In practice, the main peak should be the last peak off since there should not be any species longer than the full-length product. If peaks do appear to the right of the main peak it is indicative of residual impurities and other potential contaminants.

Fig. 6. A capillary electrophoresis of an oligonucleotide synthesis. The axes are uV absorbance units (AU) versus time on the column in minutes. The main peak is clearly the dominant peak and the “shoulder” to the left is the (n-1)-mer peak. Note, there is no material detected to the right. The ratio of the main peak to the total indicates a purity greater than 95%.



The other “industry standard” for assessing quality is **mass spectrometry**. The particular variant of mass spec is **MALDI-TOF** which stands for matrix assisted laser desorption ionization- time of flight. The principle of MALDI-TOF is that the speed at which an ion moves is inversely proportional to its mass. Thus, in the ion chamber of a MALDI instrument, materials are ionized and given the same potential energy, eV, where V is the potential and e refers to the number of charges on the ion. As the ions emerge from the ion source the potential is converted to kinetic energy in the moving ion. Ions in motion obey the rule $E = 1/2mv^2$ where m is the mass of the ion, v is velocity, and E is energy. Solving for m yields $m = 2E/v^2$. Since the amount of energy is a constant in the

MALDI, the mass of the ion can be determined by velocity alone. Velocity is simply time over distance such that the time of arrival of the ion in the detector of the MALDI; i.e., time of flight, is directly converted to velocity because the distance is a constant as well. For the application in oligonucleotide synthesis QC, MALDI-TOF has the added advantage that the mass of an oligonucleotide can be estimated with precision since the mass of the individual components (the nucleotides) is fixed. For any given sequence the expected arrival time will be given by the expected mass and any deviation from that time of arrival will indicate a deviation from ideal size and/or purity.

Degenerate Sequences and Non-Standard Bases

There will be times when you will not know the exact DNA sequence for a PCR target. Such situations arise, for example, when using sequence from one species to amplify a homologous region or gene in another species. In these circumstances, you will need to make primers that are more flexible in their specificity in order to amplify a product. This where the **degenerate primer** comes in. There are a few simple guidelines that you can use to increase the chances that you will get an appropriate amplicon. These guidelines are outlined below.

First, if you are interested in amplifying a gene homolog, you should select a region in which an amino acid sequence is conserved in as many species as you can find for your potential target. While there is no hard and fast rule, it is best to have such a conserved region run for at least six to eight amino acids. This will give you more choices as to which part to use. It is also good if as many amino acids as possible in the run have one or two codons rather than four or six. Table 1 below lists the codon specifications for the twenty amino acids:

Table 1
Distribution of amino acids by Codon Specificity

<u>One Codon</u>	<u>Two Codons</u>	<u>Three Codons</u>	<u>Four Codons</u>	<u>Six Codons</u>
Met(M)	Cys(C)	Ile(I)	Ala(A)	Leu(L)
Trp(W)	Asp(D)		Gly(G)	Arg(R)
	Glu(E)		Pro(P)	Ser(S)
	Phe(F)		Thr(T)	
	His(H)		Val(V)	
	Lys(K)			
	Asn(N)			
	Gln(Q)			
	Tyr(Y)			

As an example, consider the following amino acid sequence; GYPVVTCQWD. Using the standard nucleotide coding system: A,C,G,T; R(G or A); Y (T or C); K (G or T); M (A or C); S (G or C); W (A or T); B (G,T,C); D (G,A,T); H (A,C,T); V (G,C,A); N (all), all possible DNA sequences encoding this peptide can be specified:

Gly Tyr Pro Val Val Thr Cys Gln Trp Asp
GGN TAY CCN GTN GTN ACN TGY CAR TGG GAY

The total number of possible degenerate primers for this peptide is 16,384! There are ways to make this more manageable. You can begin by specifying a fixed 3' end. The last six nucleotides are TGGGAY. Drop off the Y and the last five nucleotides are fixed as TGGGA. The next step would be to consult a Codon Usage Table for the species whose DNA is the target. Some species have a marked preference for certain codons. The best source for such data is the Codon Usage Database (<http://www.kazusa.or.jp/codon/>). There are thousands of species listed (some more complete than others, of course) by their Genus and Species names (you must know this to use the database).

An alternative means of designing degenerate primers that can be used along with or instead of the mixed base sites shown above is the **universal base** approach. Universal bases are analog compounds that can replace any of the four DNA bases without destabilizing base-pair interactions (Loakes et al., 1995). Two universal bases commonly used are 3-nitropyrrole and 5-nitroindole (figure 7). The first universal base was 2'-deoxyinosine. It is still used extensively today but does display a slight bias in nucleotide hybridization with dI:dC being favored over other pairings (cf., Kawase et al., 1986). In the past few years truly universal base analogs have been engineered that have no pairing bias and do not alter stability (cf., Loakes, 2001 for a review and Additional Readings for design and synthesis guidelines for both universal and mixed bases).

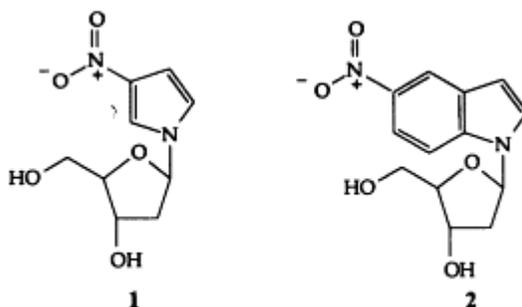


Fig. 7. Structure of the universal bases 3-nitropyrrole (1) and 5-nitroindole (2).

A new tool recently added to the array of non-standard bases in oligonucleotides that can be used for a variety of applications is the **locked nucleic acid**, or **LNA** (see Technical Report: Locked Nucleic Acids).

Summary

The history of chemical synthesis of oligonucleotides goes back to the discovery of DNA itself. The model of contemporary oligonucleotide synthesis; **protected synthesis units, solid support anchors, and end-to-end synthesis**, dates to the first attempts to synthesize polypeptides in the 1930's. The advent of phosphoramidite chemistry involving tetrazole intermediates in the 1980's made it possible to make longer and longer oligonucleotides and to make standard lengths faster and far more efficiently. Coupled with this, the transition of techniques like PAGE, CE, HPLC, and MALDI from individual lab-based research tools to high throughput analytical platforms, has made the synthesis of high quality, high purity oligonucleotides on an industrial scale an every day reality at Integrated DNA Technologies.

References

- Agarwal KL, H Buchi, MH Caruthers, N Gupta, HG Khorana, K Kleppe, A Kumar, E Ohtsuka, UL Rajbhandary, JH Van de Sande, V Sgaramella, H Weber, and T Yamada 1970 Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature* **227**: 27-34.
- Berner S, K Muhlegger, and H Seliger 1989 Studies on the role of tetrazole in the activation of phosphoramidites. *Nucleic Acids Res* **17**: 853-864.
- Hecker KH, and RL Rill 1998 Error analysis of chemically synthesized polynucleotides. *BioTechniques* **24**: 256-260.
- Kawase Y, S Iwai, H Inoue, K Miura, and E Ohtsuka 1986 Studies on nucleic acid interactions. I. Stabilities of mini-duplexes (dG2A4XA4G2-dC2T4YT4C2) and self-complementary d(GGGAAXYTTCCC) containing deoxyinosine and other mismatched bases. *Nucleic Acids Res* **19**: 7727-7736.
- Loakes D 2001 The applications of universal DNA base analogs. *Nucleic Acids Res* **29**: 2437-2447.
- Loakes D, DM Brown, S Linde, and F Hill 1995 3-nitroindole and 5-nitroindole as universal bases in primers for DNA sequencing and PCR. *Nucleic Acids Res* **23**: 2361-2366.
- McBride LJ and MH Caruthers 1983 An investigation of several deoxynucleoside phosphoramidites useful for synthesizing deoxyoligonucleotides. *Tetrahedron Lett.* **24**:245-248.
- Temsamani J, M Kubert, and S Agarwal 1995 Sequence identity of the n-1 product of a synthetic oligonucleotide. *Nucleic Acids Res* **23**: 1841-1844.

Watson JD, and FHC Crick 1953 Molecular structure for deoxyribose nucleic acid.
Nature **171**: 737-738.