Design of a novel polydactyl zinc finger with binding specificity for the proximal promoter region of PLN in the proposed treatment of cardiomyopathy

Sandy Lai and Dean Elliot

This manuscript was prepared for MCB*4050 Protein and Nucleic Acid Structure under the supervision of Prof. John Dawson, Department of Molecular and Cellular Biology, College of Biological Science.

Heart disease is the leading cause of death worldwide. Cardiomyopathy, a form of heart disease, is one of the many areas of clinical interest and is characterized as a disease of the heart myocardium. Phospholamban (PLN) is one of the many proteins linked with this disease. Its primary function in the cell is the inhibition of the sarcoplasmic reticulum Ca\(^{2+}\) ATPase (SERCA2) calcium transporter located in muscle tissue. There is a direct causal link between when PLN is over-expressed, super-inhibitory or non-existent and the development of various forms of cardiomyopathy. In an effort to counteract the development of cardiomyopathy, we have designed a novel zinc finger protein named Dandy, which specifically binds from the –77 to -95 region of the proximal promoter of the PLN gene. This protein is expected to transiently block the binding of the NFY promoter and thus, the initiation of transcription. Blocking initiation of transcription should result in an overall decrease in the concentration of PLN. Dandy is a zinc finger protein made up of two three-fingered peptides. It recognizes a 18 bp sequence with a 1 bp gap in the middle. Tertiary structure predictions of Dandy show that the protein has the intended structure of the design. More clinical research into the specific binding interactions will need to be completed before in vivo function and overall efficacy of Dandy is elucidated.

Heart failure is among the leading causes of death worldwide.\(^1\) Around 4.7 million Americans are affected, leading to approximately 17.8 billion dollars in health care costs annually.\(^1\) Current medical treatments include β-adrenergic receptor obstruction and angiotensin-converting enzyme inhibitors. Despite these however, once diagnosed, the prognosis remains rather grim, with a five-year mortality rate hovering around 50%\(^.\)\(^2\)

Cardiomyopathy is a disorder which leads to heart failure in many affected individuals. A somewhat vague term, cardiomyopathy fundamentally means a heart muscle disease or disorder. It can be caused by any number of aberrant proteins involved in regular cardiac muscle function.\(^3\) Clinically, cardiomyopathies are diagnosed as cardiac dilation and reduced cardiac systolic function resulting in dyspnea, lethargy, cyanosis and heart failure. Around 18 genes potentially play a role in the development of cardiomyopathy.\(^4\) One such gene encodes for the transmembrane phosphoprotein phospholamban (PLN). PLN is responsible for the inhibition of calcium reuptake by sarcoplasmic reticulum Ca\(^{2+}\) ATPase (SERCA2a).\(^4\)

The sarcoplasmic reticulum is one component in the contractile unit of muscle tissue. This contractile unit is known as the sarcomere.\(^5\) Within the sarcomere are bundles of myofilaments composed of actin and myosin. These proteins are arranged in a regular pattern with the actin, in conjunction with troponin and tropomyosin, forming thin filaments, which are connected to the Z line (Figure 1). When the muscle is relaxed, tropomyosin is in a position that physically prevents myosin from binding with actin. The thick filaments are formed from a coiled homodimer of myosin, which connects to the Z line via a spring-like protein called titin. Wrapped around this cylindrical unit, at the Z line, are T tubules which contain dihydropyridine receptors. Stretched between these T tubules is the sarcoplasmic reticulum. A membrane called the sarcolemma surrounds the sarcoplasmic reticulum lumen. The sarcolemma contains SERCA2, PLN and ryanodine receptors. It functions to sequester Ca\(^{2+}\) from the thick and thin filaments. During regular muscle contraction, a neuronal signal causes a depolarization of the sarcolemma. This is detected by a tetramer of dihydropyridine receptors. These receptors pull on the foot of a ryanodine receptor, causing it to open a channel and allow Ca\(^{2+}\) to flood the sarcoplasm. This free Ca\(^{2+}\) binds to troponin, which shifts tropomyosin out of the way, allowing the formation of a bond between the thick and...
Phospholamban:
Phospholamban is a 52 amino acid protein, which is subdivided into three domains,\textsuperscript{6} (Figure 2). The first domain, denoted Ia, consists of amino acids one through twenty and is the major cytosolic component of PLN. Cytosolic interaction of phospholamban with SERCA2 causes the affinity of Ca\textsuperscript{2+}-SERCA2 binding to decrease, thereby causing the inhibition of SERCA2. Section Ia contains three sites of phosphorylation. Phosphorylation of PLN causes a conformational change that alleviates SERCA2 inhibition. PLN is readily phosphorylated in three distinct sites: at serine 10 by protein kinase C (PKC), at serine 16 by cAMP activated protein kinase A (PKA), or at threonine 17 by calmodulin dependant protein kinase (CaMK).\textsuperscript{1,7} The second domain of PLN, Ib, extends from amino acids twenty to thirty and is primarily a linker between the cytosolic and transmembrane domains. The third domain, II, extends from amino acids thirty to fifty-two and is the transmembrane domain. The transmembrane domains of PLN and the SERCA2 family of proteins have a close interaction during binding. PLN is readily phosphorylated in three distinct sites: at serine 10 by protein kinase C (PKC), at serine 16 by cAMP activated protein kinase A (PKA), or at threonine 17 by calmodulin dependant protein kinase (CaMK).\textsuperscript{1,7} The second domain of PLN, Ib, extends from amino acids twenty to thirty and is primarily a linker between the cytosolic and transmembrane domains. The transmembrane domains of PLN and the SERCA family of proteins have a close interaction with each other during binding.\textsuperscript{6}

The mechanism of PLN-mediated inhibition of SERCA2 is relatively well elucidated. Initially, ATP and Ca\textsuperscript{2+} bind SERCA2 independently, forming two distinct conformations.\textsuperscript{6} One conformation has two Ca\textsuperscript{2+} bound to SERCA2, and the other has two Ca\textsuperscript{2+} and an ATP bound to SERCA2. This second conformation is a high-energy phosphoenzyme. A low-energy phosphoenzyme is formed upon transport and subsequent release of the Ca\textsuperscript{2+} into the sarcoplasmic reticulum lumen. This final stage is coupled with the hydrolysis of ATP and the release of inorganic phosphate. It is at this stage that PLN inhibits SERCA2 by binding, and forces it to remain in this conformation, rendering it unable to bind more Ca\textsuperscript{2+}. While in this conformation a groove is formed near the surface of SERCA2's transmembrane domain. PLN sits within this groove, interacting with transmembrane helices 2, 4, 6 and 9. Upon phosphorylation of PLN, there is a shift in the position of helix 2 which results in the dissociation of the PLN-SERCA2 complex.

The PLN gene is found on chromosome 6, section 6q22.1 of the human genom.\textsuperscript{4} Gene expression of PLN is varied across the body, with some tissues having an abundant supply, and virtually non-existent in others. There is even a marked difference in expression across the human heart, with higher expression in the left ventricles. It is also important to note that research done by C.F. McTiernan \textit{et. al.} indicates that there is up to 30% less PLN gene expression in the left ventricle of a failing human heart compared to a normal heart.\textsuperscript{5}

The PLN gene of a human closely resembles that found in other animals: chicken, rat, rabbit and mouse. Not surprisingly then, there is a high conservation of the sequence for –217bp upstream of the transcriptional start in other mammalian species.\textsuperscript{4} This –217bp conserved sequence contains many promoter motifs, most notably the CPI/NFY binding site, or CCAAT box, at –80 to –84. NFY is the most powerful transcriptional activator of the PLN gene.\textsuperscript{4}

\textbf{Phospholamban and Cardiomyopathy:}
Mutation or ablation of PLN can result in cardiomyopathy in a number of different ways. The basic premise is that a mutation to PLN can cause the formation of a super-inhibitor or result in a null phenotype.

There are four different naturally occurring mutations in the PLN gene that may result in some form of cardiomyopathy. The first mutation is not in the coding sequence, but rather in the promoter region of the PLN gene. A single nucleotide transition at –77bp upstream of the transcriptional start site, from A to G, has been implicated in a 1.5 fold increase of transcriptional activity relative to the wild-type.\textsuperscript{3} This mutation occurs within close proximity to the CCAAT box, which is the binding site for NFY, a regulator of PLN promoter activity.\textsuperscript{3} In other studies, it was determined that NFY requires other specific nucleotides (C/G, A/G, C/A, C/G) on the 3' side of the CCAAT box for optimal binding. This indicates that the A to G substitution at

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image.png}
\caption{The major contractile unit of muscle tissue is called the sarcomere. Influx of Ca\textsuperscript{2+} from the sarcoplasmic reticulum (not shown), which wraps around the sarcomere, causes muscle contraction via ratcheting like movement between thick and thin filaments. Adapted from SlidingMyofibril.svg found at http://en.wikipedia.org/wiki/Myofibril.}
\end{figure}
-77bp actually increases promoter binding.

The next prospective, cardiomyopathy generating mutation is in the PLN coding region, known as Arg14Del. A person afflicted with this deletion suffers from inherited dilated cardiomyopathy, often resulting in premature death. PLN was sequenced in 1203 individuals with dilated cardiomyopathy, and one was found to have a deletion of nucleic acids 39, 40 and 42. The consequence of these nucleic acid deletions was the removal of arginine 14. A study using HEK-239 cells showed that even after phosphorylation by PKA, PLN remained at least partly bound to SERCA2. In other words, the Arg14Del mutation in PLN results in a super-inhibitory form which cannot be fully reversed by phosphorylation.

Another mutation to the coding region of PLN is PLN-R9C. As mentioned earlier, one of the residues that is phosphorylated to relieve SERCA2 inhibition is serine 10. Although this mutation results in a pseudo-functional PLN, studies in transgenic mice have shown that the rate of phosphorylation with the PLN-R9C mutation is severely diminished.

The final characterized, naturally occurring mutation to the PLN coding region is a point mutation prompting the formation of a stop codon. This mutation is called T116G and occurs at the leucine 39 position. Interestingly, this mutation essentially results in a null phenotype with greater than 50% reduction in transcription (mRNA) and no functional protein. This shows some dichotomy between mouse and human models, as cardiomyopathy in mice can be alleviated by ablation of PLN, whereas its absence in humans still results in fatality due to heart failure.

This nuance between human and mouse models is important to consider when determining novel approaches to cardiac treatment. It is therefore essential to consider the intrinsic divergences between human and animal models. Some differences, such as size and lifespan are relatively apparent. However, it is also important to establish the differences in pharmacokinetics and physiology. For instance, the mouse heart beats nearly ten times faster than a human’s. More importantly, the mouse almost relies solely on SERCA2 for Ca\textsuperscript{2+} reuptake (92%), whereas it plays a smaller role (~70%) in humans.

**Zinc Fingers:**
Zinc finger proteins have a rather unique ability in that they are able to specifically bind to DNA, and sometimes RNA or proteins. Many transcription factors and DNA regulatory proteins contain a zinc finger motif. Zinc finger proteins have a largely conserved structure, usually being comprised of around thirty amino acids. It is made up of two antiparallel \( \beta \) strands and a \( \alpha \) helix. The two \( \beta \) strands and \( \alpha \) helix adopt a \( \beta\beta\alpha \) fold conformation in which the stability is highly dependent on the binding of the chelated zinc ion.

Zinc finger proteins are relatively small and do not form the same kind of hydrophobic core as larger proteins. This is why the zinc ion is generally thought to be critical to zinc finger structure. The zinc-induced stability of this structure comes from tetrahedral coordination between conserved cysteines and histidines with the zinc ion. One of cysteines is located in the \( \beta \) strand, while the other is in the linker between the \( \beta \) strands. Both histidines are located in the \( \alpha \) helix. While there is no traditional hydrophobic core, favourable interactions still occur between conserved hydrophobic residues in the strands and helix. There are three main classes of zinc finger: the Cys\textsubscript{6}-zinc cluster found in...
metabolic regulators of fungi, Cys$_2$His$_2$ found in nuclear steroid or hormone receptors and the classical zinc finger domain Cys$_2$His$_2$. It is this Cys$_2$His$_2$ zinc finger that is most commonly used in protein engineering.

Cys$_2$His$_2$ is the most common zinc finger motif in the human genome, with about seven hundred coding genes. Dimerization of zinc fingers may also occur, as is the case of the glucocorticoid receptor. Cys$_2$His$_2$ type zinc fingers tend to have a highly conserved sequence of (Tyr,Phe)-X-Cys-(Tyr,Phe)-X-Cys as confirmed by NMR structural analysis. Structural analysis also revealed that Cys$_2$His$_2$ zinc fingers tend to bind three contiguous base pairs at –1, 3 and 6 relative to the start of the α helix. The orientation of the zinc finger with respect to its 3 bp target is: –1 interacting with the 3′ region, 3 interacting in the middle and 6 interacting 5′. There may also be notable cross-strand interaction at position 2 as in when an aspartate residue acts as a hydrogen bond acceptor from the amine on either cysteine or adenine. In the three-fingered polydactyl zinc finger protein, Zif268, fingers 1 and 3 both recognize GCG. They each have a glutamic acid at position 3. The Cy and Cδ of the glutamic acids make hydrophobic contacts with the cysteine in the middle of the two target subsites. These glutamic acids also hydrogen bond with the DNA backbone amine of the residue at position –1. Contact between the zinc finger protein and cysteine or thymine are generally less favourable, especially on the 5′ end of the triplet, due to decreased potential of hydrogen bond formation and the sheer distance between the protein and DNA.

**Engineering Zinc Fingers:**

Many different methods for engineering polydactyl zinc fingers have been described. Among them are phage display, rational design and the GeneGrip technique. The method of phage display, which tests protein-DNA interaction through the amalgamation of multiple genes from a library, has a general problem with being limited by size. This problem has largely been overcome through the evolution of protein engineering techniques to assemble zinc fingers so that they recognise larger segments of DNA. The rational design approach is used in the theoretical production of zinc finger proteins either de novo or by the modification of existing zinc finger proteins. The rational design approach most commonly uses nondegenerate DNA recognition rules (Figure 3). Degenerate rules can be applied but prove to be more difficult. Nondegenerate DNA recognition rules are displayed in a simple chart, which correlates the amino acids at the –1, 2, 3 and 6 positions of the α helix to the bases found in the target DNA sequence. The GeneGrip method is reported to function in a different manner. The first difference is that the selection is completed in vivo which is thought to confer greater selectivity due to the innate differences between chromatin material encountered in the cell and in vitro DNA. Another change is that the zinc fingers are derived from intact wild type zinc fingers which bring an added benefit of being less immunogenic. Lastly, the GeneGrip shuffling approach allows for simultaneous production of thousands of zinc finger proteins.

An important aspect of engineering a zinc finger is the uniqueness or specificity of its binding. To ensure selectivity of a zinc finger protein, a 16 bp sequence, which statistically only occurs every 4.3 billion nucleotides, or an 18 bp sequence, which occurs once every 68 billion nucleotides, should be chosen.

There are three methods that are used in polydactyl zinc finger design: parallel, sequential and bipartite selection. Parallel selection assumes that there is no loss of function upon combining zinc fingers. A loss of function can sometimes occur when the recognition head of the α helix interacts with four, as opposed to three, base pairs causing an overlap. This extra interaction occurs at the +2 region and is a cross-strand interaction. This problem can be circumvented by the amalgamation of multiple genes from a library, which is tests protein-DNA interaction through the evolution of protein engineering techniques to assemble zinc fingers so that they recognise larger segments of DNA. The rational design approach is used in the theoretical production of zinc finger proteins either de novo or by the modification of existing zinc finger proteins. The rational design approach most commonly uses nondegenerate DNA recognition rules (Figure 3). Degenerate rules can be applied but prove to be more difficult. Nondegenerate DNA recognition rules are displayed in a simple chart, which correlates the amino acids at the –1, 2, 3 and 6 positions of the α helix to the bases found in the target DNA sequence. The GeneGrip method is reported to function in a different manner. The first difference is that the selection is completed in vivo which is thought to confer greater selectivity due to the innate differences between chromatin material encountered in the cell and in vitro DNA. Another change is that the zinc fingers are derived from intact wild type zinc fingers which bring an added benefit of being less immunogenic. Lastly, the GeneGrip shuffling approach allows for simultaneous production of thousands of zinc finger proteins.

Another change is that the zinc fingers are derived from intact wild type zinc fingers which bring an added benefit of being less immunogenic. Lastly, the GeneGrip shuffling approach allows for simultaneous production of thousands of zinc finger proteins.

**Figure 4:** Target DNA sequence of polydactyl Dandy protein located at -95 to -77 of human PLN’s proximal promoter. Dandy binds an 18 bp sequence. A 1 bp gap is found at -86 shown in black. The order of the zinc fingers, which bind to the sequence, is also given.

**Figure 5:** Designated amino acid residues for each DNA binding position (-1, 3, 6 of the recognition helix.) The following recognition codes were used in this alignment: i) nondegenerate code table [16], ii) Choo and Klug’s phage display results [26], iii) Segal’s 5’-GNN-3’ results [27], iv) Dreier’s 5’-GNN-3’ results [28], v) Dreier’s 5’-CNN-3’ results [24], vi) Dreier’s 5’-ANN-3’ results [23], and vii) Zinc Finger Tools application, http://www.scripps.edu/mb/barbas/zfdesign/zfdesignhome.php [21]
through carefully picking the recognition sequence i.e. using 5'-GNN-3' domains. If target site overlap is avoided, this method would allow for the protein to be designed modularly, that is, each finger is designed individually. Sequential selection is much like it sounds, in that the preceding zinc fingers are used to determine the arrangement of following zinc fingers.10,20 This technique avoids the problem of overlap; however, since it requires the production of six libraries for each finger, it is often beyond the reach of many laboratories. The bipartite selection method was more recently developed and attempts to combine the advantages of the parallel and sequential selection methods.19 It is successful in that it avoids target site overlap and only requires two zinc finger protein libraries, thus decreasing the time and complexity of the sequential method. However, it is limited in that it requires multiple rounds of phage display.

**Conclusion / Aim of Research:**

The overall goal of this line of research is to design a novel six-fingered polydactyl zinc finger protein with specific DNA interaction to the sequence 5'-ACTGTGGCTAACCAATCGA-3'. This sequence is located at −77 to −95, in the promoter region of the PLN gene. It is intended to transiently block the NFY promoter from binding to the CCAAT site, located at −80 to −84. In this way, transcription of the PLN gene may be blocked wherever the ZFP is bound. This poses possible health benefits and additional treatment options to people affected with cardiomyopathy.

**MATERIALS AND METHODS**

The target DNA sequence of the newly designed polydactyl zinc finger protein, named Dandy, is 18 bp long with a 1 bp gap in the middle (Figure 4). The gap was placed within the target sequence to accommodate the addition of a longer linker, which was used to join the two 3-fingered peptides together. Dandy’s target sequence was specifically chosen to bind within PLN’s proximal promoter from −77 to −95 to disrupt the binding of the transcriptional promotor NFY. The Cp1/NFY element has a consensus sequence of CCAAT that is conserved in mammalian phospholamban genes (-84 to -80 in humans). This element has been reported to be absolutely necessary for transcription in rats.1 Considering all the proteins in eukaryotic cells that interact with the CCAAT box, NFY is almost exclusively the only one requiring all five nucleotides for binding.5 Since NFY and Dandy compete for the same target sequence, when Dandy is bound, the NFY promoter cannot bind. Thus it will be unable to recruit the transcriptional machinery necessary for PLN transcription to occur and result in an overall down regulation in PLN expression.

Other considerations were taken into account in choosing this sequence. Since no recognition code based on 5'-TNN-3' triplets exist, the chosen sequence contains no triplets that begin with thymine. There is also no target site overlap in the sequence. We want this because the protein is designed modularly, thus 3 bp sites are desired over 4 bp recognition subsites so we can assemble each finger independently.20 Target site overlaps have been found in 5'-GNG-3' triplets accompanied by an A or C at the 3' end.21 The target sequence contains no such triplets. The 1 bp gap between the two 9 bp sub-target sequences is due to a longer linker between the two three-fingered peptides of Dandy. This puts less strain on the DNA.22 The recognition site for fingers 4, 5 and 6 of the designed protein contains 3 guanosine residues as per the nondegenerate code rules16 which should ensure tight binding. Recognition sites for fingers 1, 2 and 3 contain only one guanosine, and thus the fingers were designed based on zinc finger domain sequences shown to be specific for 5'-ANN-3'16,23 and 5'-CNN triplets.24

When designing our polydactyl zinc finger protein some basic criteria were applied. To ensure that the protein is a novel design, it must not contain two consecutive fingers with greater than 60% identity and must not have greater than 85% homology with any other protein. Residues in the recognition portion of the α helix, the ones involved in DNA binding, should be chosen using the best possible agreement among the different recognition codes. It is also of great importance that each finger match its consensus sequence which follows the general guide: (Y,F)-X-C-X2-5-C-X3-(Y,F)-X5-L-X2-H-X3-5-H consensus sequence.25

Their results showed highly variable residue positions. The results that occurred most often were used in our alignment. There are also codes available for the recognition of 5'-ANN-3';27,28 5'-CNN-3';24 and 5'-ANN-3' triplets.21 These were also considered in our design. In addition, our target sequence was inputted into the Zinc Finger Tools application23 and the binding residues given were added to the alignment. After a comparison of results, the most common residues were chosen. These turned out to be the exact ones given by the Zinc Finger Tools application. Because of this, the residues that the application gave for the rest of the recognition helix were incorporated into the designed protein. It should be noted that residue 4 is a leucine according to the (Y,F)-X-C-X2-5-C-X3-(Y,F)-X5-L-X2-H-X3-5-H consensus sequence.12
However, the protein backbone given by the Zinc Finger Tools application was not used but was designed separately. The rest of the finger was based on the Sp1C framework. The N- and C-terminal backbone of each finger is based on that of Sp1, YKCPECGKSFS and HQRTH, respectively. However, some conservative amino acid substitutions were made, such as R for K and Y for F. This was done to lower the number of identities between the designed proteins and other existing proteins. This backbone contains the conserved cysteine and histidine residues that bind the zinc ion. It is also consistent with the consensus sequence mentioned above.

The individual fingers were then joined together by linker sequences. The six-fingered protein is actually composed of two three-fingered peptides joined together. Canonical linkers join the fingers in each of the three-finger peptides. Canonical linker, also called a Kruppel linker, is a short sequence consisting of TGKEK. This sequence does more than link the fingers together. It also caps the recognition helix at its C-terminal which helps with its stabilization. While these short linkers are suitable for the three-fingered peptides, they should not link every single finger together. Studies have shown that using only canonical linkers in six-fingered proteins resulted in proteins with no improvement in binding affinity over three-fingered proteins. This is due to strain put on the DNA. When a zinc finger binds to DNA, the helix is slightly unwound. While favourable docking for three-fingered proteins can be accomplished using canonical linkers, additional fingers require a longer linker to prevent serious strain on the DNA. Thus, a longer linker was used to join the two 3-fingered peptides together. The RQGDGERP linker from Kim and Pabo’s experiment was used. This results in a 1 bp gap in the target sequence.

The N terminus for the whole protein was also from Sp1 (LEPGEK). The C-terminus from Sp1 (TGKKTS) was added onto the end of finger 6. This is followed by the nuclear localization signal, PPKKKRKV. This signal is recognized by nuclear import receptors, which allows for the protein to be transported into the nucleus. As in other nuclear proteins, the nuclear localization signal is located on the C terminus. The complete sequence of the designed protein can be seen in Figure 6. Figure 7 shows the docking arrangement of Dandy to its target sequence.

**Results and Discussion**

**Target sequence:**

Since there is a one base pair gap between the recognition sites for the two 3-fingered peptides, the zinc finger protein can bind to four possible sequences (with A, C, G, or T in the gap.) BLAST similarity searches performed with all four sequences show high homology with other human chromosomal sequences (Figure 8). However, only the human phospholamban promoter contained both recognition sites in full. Thus, the designed protein can only bind to this sequence with high specificity.

The predicted structure of the target sequence is B-DNA (Figure 9). This is the form of DNA that is present under *in vivo* conditions. It is a right-handed helix and is fully hydrated. There are 10 bp per turn. Base pairing between nucleotides will result in the formation of a minor groove and a major groove within the DNA helix. It is in the major groove that the zinc finger binds. As previously mentioned, this binding also results in a slight unwinding of the B-DNA, causing some strain in the structure. In the nucleus, DNA is present as chromatin, a tightly packed structure. This presents an obstacle for protein binding as the desired gene may be wound up inside the structure. However, Dandy should still be able gain access to its target site since the target is close to other known transcriptional regulators. Since other transcriptional regulators can access that portion of the PLN gene, it is an indication that Dandy would be free to do so as well.
Target sequence: ACTGTGGCTAACCAATCGA 19 bp

<table>
<thead>
<tr>
<th>Definition</th>
<th>E value</th>
<th>Bit Score</th>
<th>Identities</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human DNA sequence from clone RP3-509L4 on chromosome 6q22.1-22.33 contains PLN gene</td>
<td>0.12</td>
<td>38.2</td>
<td>19</td>
<td>0%</td>
</tr>
<tr>
<td>Homo sapiens phospholamban gene, proximal promoter and exon 1</td>
<td>0.12</td>
<td>38.2</td>
<td>19</td>
<td>0%</td>
</tr>
<tr>
<td>Homo sapiens chromosome 8, clone RP11-383H13, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
</tbody>
</table>

Possible target sequence: ACTGTGGCTACCAATCGA 19 bp

<table>
<thead>
<tr>
<th>Definition</th>
<th>E value</th>
<th>Bit Score</th>
<th>Identities</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens chromosome X clone Qc-7G6, QLL-F1720, QLL-C1335, Qc-8B7, Qc-11H12, Qc-7F6, QLL-E153, Qc-10E8, Qc-10B7 map q28, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
<tr>
<td>Homo sapiens BAC clone RP11-432019 from 7, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
<tr>
<td>Homo sapiens chromosome X clone QLL-D9139, Qc-7G12, Qc-7C1, Qc-12B2, Qc-12D5, QLL-A074, Qc-9D3 map q28, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
</tbody>
</table>

Possible target sequence: ACTGTGGCTGACCAATCGA 19 bp

<table>
<thead>
<tr>
<th>Definition</th>
<th>E value</th>
<th>Bit Score</th>
<th>Identities</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens chromosome 3 clone RP11-180N14, complete sequence, length=169709</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
<tr>
<td>Homo sapiens chromosome 7 clone RP11-371N6, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
<tr>
<td>Homo sapiens chromosome 7 clone RP11-140I14, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
</tbody>
</table>

Possible target sequence: ACTGTGGCTTACCAATCGA 19 bp

<table>
<thead>
<tr>
<th>Definition</th>
<th>E value</th>
<th>Bit Score</th>
<th>Identities</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens chromosome 3 clone RP11-24F11, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
<tr>
<td>Human DNA sequence from clone RP3-459L4 on chromosome 6p22.3-24.1 contains part of the SCA1 gene for spinocerebellar ataxia 1 (olivopontocerebellar ataxia 1, autosomal dominant, ataxin 1), complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
<tr>
<td>Homo sapiens ccr2b (ccr2), ccr2a (ccr2), ccr5 (ccr5) and ccr6 (ccr6) genes, complete cds, and lactoferrin (lactoferrin) gene, partial cds, complete sequence</td>
<td>7.4</td>
<td>32.2</td>
<td>16</td>
<td>0%</td>
</tr>
</tbody>
</table>

Figure 8: Top three similarity search results for each of the four possible target DNA sequences with other human chromosomal sequences acquired from a BLAST search from NCBI http://www.ncbi.nlm.nih.gov/BLAST/ [35]
Figure 9: Tertiary structure of the target sequence as obtained from the model.it server http://hydra.icgeb.trieste.it/dna/model_it.html [36]

<table>
<thead>
<tr>
<th>Definition</th>
<th>E value</th>
<th>Bit Score</th>
<th>Identities</th>
<th>Positives</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc finger protein 572 [ZNF572] [Bos taurus (Bovine)]</td>
<td>3e-60</td>
<td>233</td>
<td>54%</td>
<td>77%</td>
<td>2%</td>
</tr>
<tr>
<td>Zinc finger protein 572 [ZNF572] [Homo sapiens (Human)]</td>
<td>1e-56</td>
<td>221</td>
<td>48%</td>
<td>65%</td>
<td>12%</td>
</tr>
<tr>
<td>Zinc finger protein 436 (Zinc finger protein 46) [Znf436] [Mus musculus (Mouse)]</td>
<td>9e-56</td>
<td>218</td>
<td>52%</td>
<td>71%</td>
<td>2%</td>
</tr>
<tr>
<td>Zinc finger protein 46 (Zfp-46) (Zinc finger protein MLZ-4) [Zfp46] [Mus musculus (Mouse)]</td>
<td>9e-56</td>
<td>218</td>
<td>52%</td>
<td>71%</td>
<td>2%</td>
</tr>
<tr>
<td>Zinc finger protein 436 [ZNF436] [Pongo pygmaeus (Orangutan)]</td>
<td>1e-55</td>
<td>218</td>
<td>52%</td>
<td>71%</td>
<td>2%</td>
</tr>
<tr>
<td>Zinc finger protein 436 [ZNF436] [Homo sapiens (Human)]</td>
<td>1e-55</td>
<td>218</td>
<td>52%</td>
<td>71%</td>
<td>2%</td>
</tr>
<tr>
<td>Zinc finger and SCAN domain-containing protein 2 (Zinc finger protein 29 homolog) (Zfp-29) [ZSCAN2] [Homo sapiens (Human)]</td>
<td>3e-55</td>
<td>216</td>
<td>50%</td>
<td>73%</td>
<td>2%</td>
</tr>
<tr>
<td>Isoform 2 of Q727L9 [ZSCAN2] [Homo sapiens (Human)]</td>
<td>3e-55</td>
<td>216</td>
<td>50%</td>
<td>73%</td>
<td>2%</td>
</tr>
<tr>
<td>Hypothetical protein DKF2p469D0914 [DKF2p469D0914] [Pongo pygmaeus (Orangutan)]</td>
<td>8e-55</td>
<td>215</td>
<td>50%</td>
<td>73%</td>
<td>2%</td>
</tr>
<tr>
<td>CDNA FLJ16290 fis, clone OCBBF2019684, highly similar to ZINC FINGER PROTEIN 29 [Homo sapiens (Human)]</td>
<td>1e-54</td>
<td>214</td>
<td>50%</td>
<td>73%</td>
<td>2%</td>
</tr>
</tbody>
</table>

Figure 10: Top ten similarity search results between Dandy and other naturally occurring proteins as obtained from a BLAST search from ExPASy http://expasy.org/tools/blast/ [35]
Protein homology and structure:
Dandy is most similar to other zinc finger proteins (Figure 10). The protein with highest similarity is a bovine zinc finger protein, ZNF572, with 54% identity and 77% homology. Other similar proteins include ZFPs from humans, mice, and orangutans. Since Dandy is most similar to other ZFPs, this may be an indication that the intended design would have similar functionality in DNA binding.

The secondary structure prediction methods are based on three states: helix, extended strand, and coil. Some regions of the protein predicted by GOR and SOPMA match what was intended by the design, but some portions were not (Figure 11). Overall, the SOPMA method of prediction was in closer correlation to what the protein was designed to be. It predicted more helical regions than the GOR method. The linker regions were predicted to be random coils as expected but the residues CPDC for each finger were predicted to be coils in both methods instead of a beta turn as intended.

Secondary structure prediction is difficult as it is not accurate 100% of the time. GOR was developed by Garnier et al. and is based on information theory. It correctly predicts secondary structure 64.4% of the time. This method predicts a helical segment when there are at least four residues in a row that fit that profile. Thus, if residues were predicted to be helix-helix-extended-helix-helix, the method would still not predict a helical structure. This may be why there were less helices predicted by GOR than by SOPMA. The self-optimized prediction method (SOPMA) correctly predicts 69.5% of secondary structure. This slightly higher degree of accuracy may explain why it more closely matched the design of the protein description of the protein.

It should be noted that while the secondary structure predictions do not greatly match with the intended design, the tertiary structure does match the intended design. The tertiary structure prediction of Dandy by SWISS MODEL can be seen in Figure 12. Each finger has a helix and two beta strands, as intended by the design. The fingers are linked by the canonical linkers in short loops into three-fingered peptides, the longer linker in the larger loop in turn joins these peptides.

A Ramachandran plot of our protein is shown in Figure 13. The majority of amino acids are found in the upper left quadrant and the centre/bottom left quadrant. These correspond to the β region and the α region, respectively. These groupings are expected since the zinc protein has numerous beta strands and alpha helices in its structure. There are also some amino acids in the loop region. These correspond to the residues present in the linkers, and perhaps in the nuclear localization signal. The loop region also contains any left-handed helices present in the protein. The only true outliers of the plot are glycine residues. This is because glycine has only a single H as its side chain which gives it more freedom of movement. Thus, it can adopt many more conformations than any other residue. This freedom allows it to assume phi and psi angles that other amino acids are unable to achieve.

Protein-DNA and intramolecular interactions
An arginine residue binds every guanosine residue in the target sequence. This is due to the formation of two hydrogen bonds between them. Similarly, adenine is bound by glutamine and asparagine in this way. Thymine may be bound by serine or threonine via hydrophobic interactions. The methyl group in thymine interacts with the methyl group of serine/threonine. In the designed protein, every thymine is recognized by a threonine. Cytosine is recognized by glutamate or aspartate. These amino acids may hydrogen bond with cytosine. In addition, their carboxyl group can participate in electrostatic interactions with the amino group of cytosine. These interactions can be seen in Figure 14. In addition to these, other residues play a role in DNA binding. Residues -2, 1, and 5 of the helix may make contacts with the phosphate backbone. These contacts may be important for proper docking of the protein to the DNA. The lysines in the
Figure 13: Ramachandran plot for Dandy. Crosses are the amino acids contained in Dandy’s sequence, except for glycines which are represented by squares. The majority of residues are located in the upper left hand quadrant (beta strands) and the central/bottom left quadrant (alpha helices). Only a few amino acids (besides glycine) are in the left-hand helical/loop region. Plot created from http://xray.bmc.uu.se/cgi-bin/gerard/rama_server.pl [39]

Figure 14: Schematic depiction of the interactions that occur between nucleotides and amino acids. [16]
linker region also contribute to this effect through water-mediated phosphate contacts. In addition, the leucine at residue 4 of the helix participates in hydrophobic packing in the protein core. As explained in the introduction, the conserved cysteine and histidine residues are coordinated in a tetrahedral fashion. This arrangement can be seen in Figure 15.

Conclusion:
We believe that the overall design strategy outlined in this paper is sufficient for the creation of a novel zinc finger protein. By designing the protein in a modular fashion using recognition codes, Dandy is predicted to have the correct tertiary structure and to be able to recognize the specific target sequence. This can be seen in the Figure 12, which shows that the conserved amino acid substitutions did not greatly distort the desired zinc finger protein design. Also, Figure 5 shows the large consensus between recognition codes for the residues selected for the DNA binding positions of the recognition helix. However, some potential problems should be noted. Since the protein is actually two three-fingered peptides separated by a linker longer than TGEKP, it may bind an untargeted sequence. The first three fingers of the six-fingered protein will have a higher binding affinity as a whole to the entire target sequence than the individual three-fingered peptide’s binding affinity to their 9 bp sequence. Subsequent studies would investigate this matter further to determine if the design of the protein needs to be altered.

Figure 15: Zinc ion coordination by the conserved cysteine and histidine residues. Both histidines reside on the helix. One cysteine is located in a beta strand while the other is in the turn between the beta strands.

REFERENCES