information

Article

Structural and Functional Modeling of Artificial Bioactive Proteins

Nikola Štambuk [1],* and Paško Konjevoda [2]

[1] Center for Nuclear Magnetic Resonance, Ruđer Bošković Institute, Bijenička cesta 54, 10002 Zagreb, Croatia
[2] Laboratory for Epigenomics, Division of Molecular Medicine, Ruđer Bošković Institute, Bijenička Cesta 54, 10002 Zagreb, Croatia; pkonjev@irb.hr
* Correspondence: stambuk@irb.hr; Tel.: +385-1-468-0193

Academic Editors: Abir U. Igamberdiev, Pedro C. Marijuán and Lin Bi
Received: 5 November 2016; Accepted: 1 March 2017; Published: 5 March 2017

Abstract: A total of 32 synthetic proteins designed by Michael Hecht and co-workers was investigated using standard bioinformatics tools for the structure and function modeling. The dataset consisted of 15 artificial α-proteins (Hecht_α) designed to fold into 102-residue four-helix bundles and 17 artificial six-stranded β-sheet proteins (Hecht_β). We compared the experimentally-determined properties of the sequences investigated with the results of computational methods for protein structure and bioactivity prediction. The conclusion reached is that the dataset of Michael Hecht and co-workers could be successfully used both to test current methods and to develop new ones for the characterization of artificially-designed molecules based on the specific binary patterns of amino acid polarity. The comparative investigations of the bioinformatics methods on the datasets of both de novo proteins and natural ones may lead to: (1) improvement of the existing tools for protein structure and function analysis; (2) new algorithms for the construction of de novo protein subsets; and (3) additional information on the complex natural sequence space and its relation to the individual subspaces of de novo sequences. Additional investigations on different and varied datasets are needed to confirm the general applicability of this concept.

Keywords: synthetic protein; structure prediction; function prediction; de novo design

1. Introduction

Proteins are molecules fundamental to life, and their biological function is driven by their structure [1,2]. The modeling of the protein structure and function relationship is important and challenging [1,2]. Recent progress in protein biochemistry and biophysics has enabled the construction of artificial (de novo) proteins with specific properties [1–3]. The predominant part of the possible protein sequences and structures not tested by evolution may be evaluated by de novo protein design, which could provide solutions to new protein-structure/function targets [2–6]. The goal of designing de novo proteins is to construct the molecules that structurally and functionally mimic natural proteins and to discover new structure-function relationships compared with those found in nature [2,6].

The sequence space of proteins is huge and complex [1,2]. It has evolved in time influenced by the evolutionary processes of selection and mutation [1,2]. By contrast, the subsets of de novo designed protein sequences, including the one tested in this research, are limited, internally consistent, of high sequence identity and artificially designed for a specific purpose. There is no standardized dataset of artificial proteins for the testing of bioinformatics algorithms, in contrast to the naturally-occurring proteins [7]. However, there are several sets of well-characterized artificial proteins that may be used for the testing of algorithms concerning protein structure and function [1–6]. In our research, we tested standard bioinformatics methods using the de novo protein subset of Hecht et al. [3,6,8–12]. It represents a well-characterized and sufficiently large subset of 15 synthetic α- and 17 β-proteins

(Hecht_α and Hecht_β) [3,6,8–13]. A dataset of this size was sufficient for this study due to the small number of variables analyzed in comparison to the number of members of each class (Hecht_α and Hecht_β) [14]. For the structural characterization of the dataset, Hecht and co-workers used: size-exclusion chromatography, NMR and circular dichroism spectroscopy (CD), X-ray crystallography, electrospray mass spectrometry (ESMS), differential scanning calorimetry (DSC) and several other methods [3,6,9,12,15–21].

Michael Hecht and his co-workers [8–10,22] were the first to design functional de novo protein structures, as well as simple algorithms based on binary polar(p) and nonpolar(n) amino acid patterning. They showed that complex molecular structures could be made using amino acid polarity patterns pnppnnp and pnpnpnp that define stable α-helices and β-strands, respectively [8–10]. Hecht_α proteins were not explicitly designed to be functional, but recent studies have shown that they also provide the biological functions necessary to maintain cell growth where genes encoding enzymes essential for amino acid biosynthesis have been deleted [6,11]. They have been named SynRescue proteins [3,6,11].

The existing methodology for protein structure-function analysis has been derived and tested using natural proteins, so that, as a rule, the construction patterns for different de novo protein subsets have been extracted from the natural sequence space [1,2]. Until recently, due to the small number of non-natural proteins, it was not possible to analyze or test the standard bioinformatics methods on sufficiently large datasets. The overall goal of the paper is to investigate whether the use of standard bioinformatics algorithms is applicable for efficient and accurate structure-function modeling of the synthetic proteins subsets Hecht_α and Hecht_β. The applicability of the presented methodology will be discussed considering de novo protein subsets recently reported by Woolfson and co-workers [4,5,23–25] and Baker and co-workers [1,26].

2. Results and Discussion

The de novo protein design of Hecht et al. [9,10,12] is based on the empirical observation that the second base of the nucleotide triplet of the genetic code (Table A1) specifies amino acid polarity, i.e., the second U/T of the codon specifies nonpolar amino acids in the selected protein α-strands and β-sheets, while the second A of the codon specifies polar amino acids (Section 3.1). The dataset of 32 newly-designed α- and β-proteins (Hecht_α and Hecht_β) consisting of well-defined and stable structures is the first sufficiently large subset of synthetic sequences that can be used for the testing of standard bioinformatics models and algorithms, derived from the natural protein sequences (Tables S1 and S2) [3,6,9,12,13].

First, we will investigate:

1. computational techniques to detect periodicity in Hecht_α helices and Hecht_β sheets and hydrophobicity values assigned to the individual amino acids along different protein structural segments [27,28];
2. secondary structure elements of Hecht_α and Hecht_β proteins, surface accessibility, antigenicity and solubility [29–36].

We will also analyze specific functional aspects of the artificial α-protein sequences that arise from the ligand-receptor interplay of their 3D structure and the reactive patterns of their natural ligands (heterogens) [37–39]. This second aspect of the protein structure-function relationship will be inspected utilizing:

1. the protein virtual spectroscopy technique, i.e., the informational spectrum method (ISM), based on the amino acid electron-ion interaction potential (EIIP) [40–44];
2. the 3DLigandSite method that uses predicted Hecht_α and Hecht_β protein structures and the ligands present in homologous natural structures to predict ligand binding sites [38,39].

2.1. Spectral Analysis of Hecht_α and Hecht_β Proteins

We inspected the periodicity in Hecht_α (Table S1) and Hecht_β (Table S2) protein structures using the method of Cornette et al. [27], which is based on the results of 38 published hydrophobicity scales compared for their ability to identify the characteristic 3.6 residue period of α-helices (Table A2). As suggested by the authors, we applied the normalized PRIFT scale since this technique maximizes the amphipathic index of the Fourier transform [27]. In addition to the Fourier sequence analysis, we also performed an alternative least-squares spectral analysis (LSSA) that, for short peptide components, provides a more reliable estimate of periodicity [27].

Both techniques of spectral analysis led to the same result and confirmed that one pronounced frequency peak at the position $x = 0.28$ characterizes all Hecht_α proteins (Table 1). Hecht_β proteins are characterized by another pronounced frequency peak at a different position, $x = 0.45$, which enables simple and accurate virtual spectroscopy screening of both artificial protein classes. Moreover, the peak positions of both artificial protein classes presented in Table 1 and Figure 1 are in marked agreement with previously-published results for natural proteins [27,45].

The value of 0.28 (i.e., 101°) that we measured for Hecht_α proteins is identical to the finding of Eisenberg et al. [45] that 157 segments of α-helix exhibited a peak at 100°. The α-structural repeat of 3.6 residue/turn, approximated by polar and nonpolar residue patterns pnppnnp [9], is obtained from the dominant peak value at $360°/101° \cong 3.6$ (calculated according to Cornette et al. [46]). For Hecht_β proteins, we identified one distinct frequency peak at 0.45 (i.e., 162°), which is confirmed by the maximum peak at 160° reported by Eisenberg et al. [45] for the average of 220 strands of β-structure. Therefore, it is not surprising that the methods of Eisenberg et al. [45] and Cornette et al. [27] predict the same peak positions for the tested proteins.

Figures A1 and A2 show that several de novo α-helical and β-sheet peptides reported by Woolfson and co-workers [23–25], Quinn et al. [47], Schneider et al [48], Griffioen et al. [49] and Baker and co-workers [26] exhibit almost identical frequency peaks when compared to the Hecht_α and Hecht_β dataset (Table 1, Figure 1). These distinct α-helical and β-sheet frequency peaks are within the range predicted by Eisenberg et al. [45] for α helices (mean = 0.28, range from 0.25–0.31) and β strands (mean = 0.44, range from 0.39–0.50).

Table 1. Spectral analysis of Hecht_α and Hecht_β proteins (method by Cornette et al. [27]).

Synthetic Proteins	Frequency Peak [1]	Amino Acid/No.
Hecht_α proteins (SynRescue)		
SynSer (B1 to B4)	0.28	Q58
SynGltA1	0.28	Q58
SynIlvA (1 and 2)	0.28	Q58
SynFes (1 to 8)	0.28	Q58
Hecht_β-proteins		
#4, #7, #23, #43	0.45	E57
#66, #68, #69, #75	0.45	E57
#8, #10, #12, #16	0.45	D57
#17, #19, #24, #71, #78	0.45	D57

[1] Fourier and least-squares spectral analyses.

It is worth mentioning that although large datasets of natural proteins may be used to extract characteristic peaks for α- and β-structures, the procedure may not always work well for the natural proteins on an individual basis [27,45]. The absence of distinct peaks in one natural α- and one natural β-protein (of experimentally-determined structure [13,22]) is shown in Figure A3 (proteins 1cc5 and 1amg-2-AS, respectively). By contrast, all Hecht_α and Hecht_β proteins can be predicted individually (Table 1 and Figure 1), which enables fast structural screening by means of the computational techniques presented and the testing of new bioinformatics methods and algorithms for structural predictions.

This is because distinct polarity patterns encode the well-defined artificial structure of all Hecht_α and Hecht_β proteins [3,9].

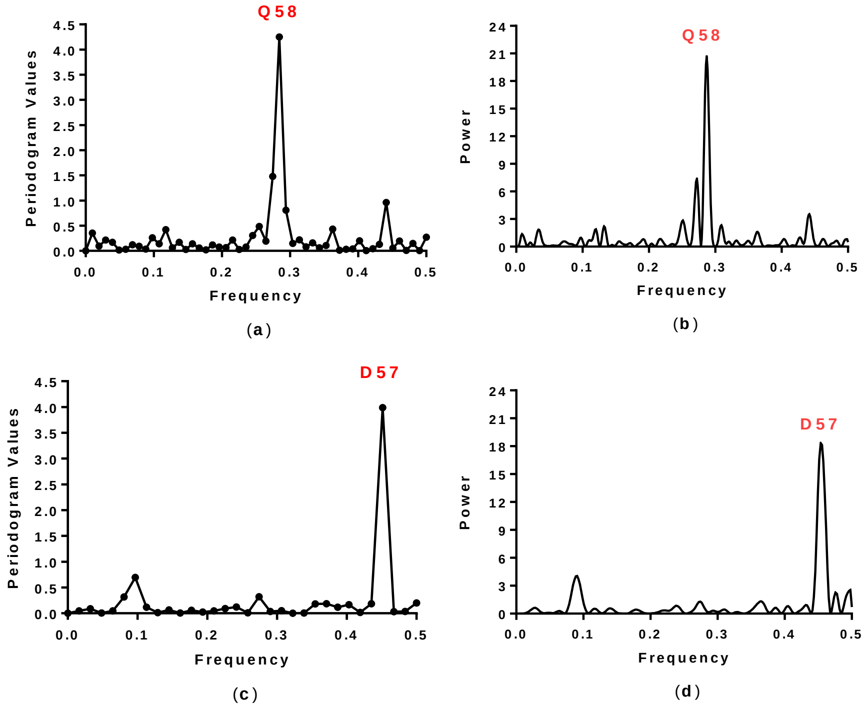


Figure 1. Characteristic peaks of Hecht_α protein SynSerB3 and Hecht_β protein #17 were determined using the method of Cornette et al. [27]. (**a**) Hecht_α protein SynSerB3 exhibits the peak at $x = 0.28$ (Fourier spectral analysis); (**b**) SynSerB3 exhibits an identical peak ($x = 0.28$) with the least-squares spectral analysis; (**c**) Hecht_β protein #17 exhibits the peak at $x = 0.45$ (Fourier spectral analysis); (**d**) #17 exhibits an identical peak ($x = 0.45$) with the least-squares spectral analysis. The amino acids belonging to the detected hot spots are marked in red.

2.2. *Hydropathy Analyses of Hecht_α and Hecht_β Proteins*

The hydropathy of Hecht_α and Hecht_β proteins was investigated using a sliding block based on the Kyte–Doolittle scale (Table A2) [50]. This method sums up the hydrophobicity values of amino acid residues. It is often used for identifying surface-exposed regions, as well as transmembrane regions, depending on the size of the sliding block used, e.g., a short window of 7–9 is used for the exposed regions and a large window of 19–21 is for the transmembrane regions [28,50,51].

Figure 2 shows the position of three predicted surface-exposed regions (P1, P2 and P3) of Hecht_α and Hecht_β proteins based on three-point moving average values of the nine amino acid sliding block.

Detected peaks of Hecht_α proteins predict combinatorial turn positions and possible antigenic sites. Figures 2a and 3a,b show that out of three predicted sites, the first and the third sites (P1 and P3) locate two important amino acids, 25 and 77, which precede residues at positions 26 and 78; these latter presumably stabilize the dimeric structure of Hecht_α SynRescue proteins with charged or hydrogen bonding groups [3]. This is valid for all 15 Hecht_α SynRescue proteins presented in Figure 2a and confirms the results of Murphy, Greisman and Hecht regarding Hecht_α-protein structure [3]. The presumed interactions between P1 and P3 regions indicate that there could be an antigenic site of SynRescue proteins accessible at the interaction-free P2 position, which is predicted by several methods to be in the vicinity of region 53–59 (Figure 3a–c), i.e., near the turn between helix 2 and helix 3 [3,6], while the amino acids at or near positions 26 and 78 could influence dynamic structures that fluctuate between monomeric and dimeric states [3]. Hydropathy analyses of Hecht_α and Hecht_β proteins based on the PRIFT scale of Cornette et al. [27] predicted the same bioactive sites as the Kyte–Doolittle scale (Figure 2).

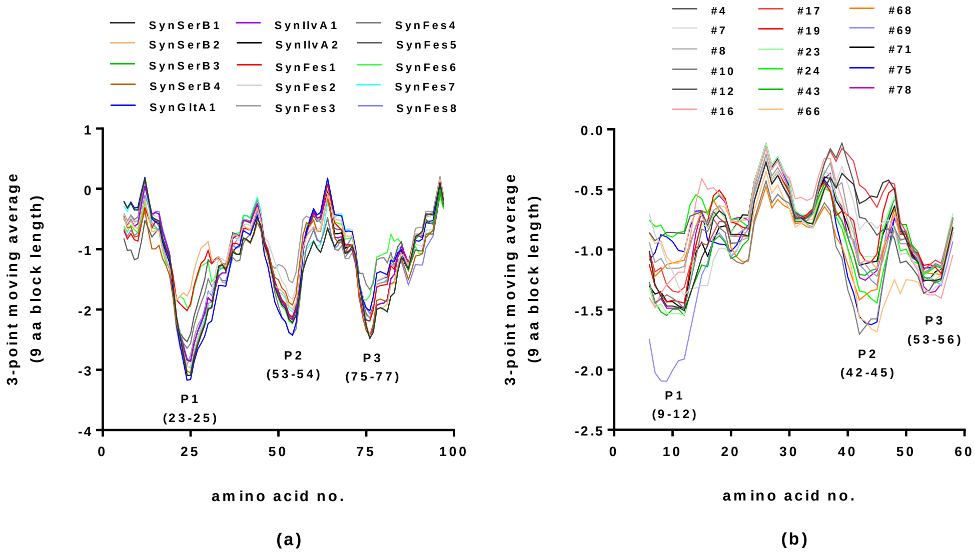
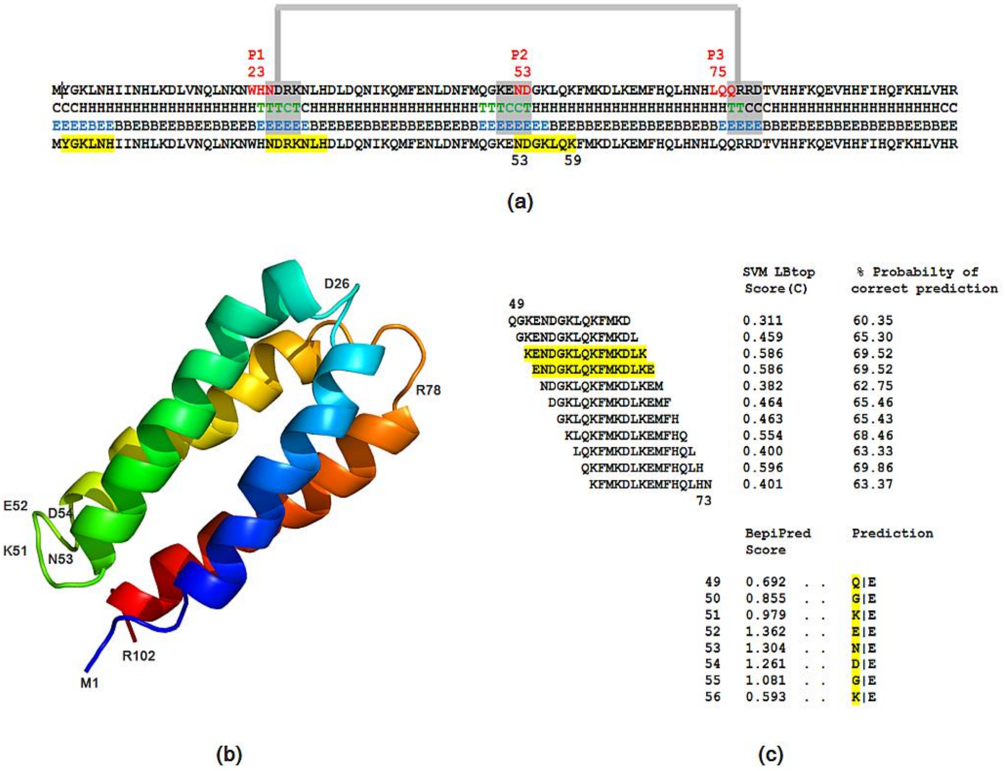


Figure 2. (**a**) Surface-exposed regions (P) of 15 Hecht_α proteins identified using the Kyte–Doolittle scale; (**b**) surface-exposed regions (P) of 17 Hecht_β proteins identified using the Kyte–Doolittle scale.



```
                     P1                                P2                         P3
                     23                                 53                         75
MYGKLNHIINHLKDLVNQLNKNWHNDRKNLHDLDQNIKQMFENLDNFMQGKENDGKLQKFMKDLKEMFHQLHNHLQQRRDTVHHFKQEVHHFIHQFKHLVHR
CCCHHHHHHHHHHHHHHHHHHHTTTCTCHHHHHHHHHHHHHHHHHHHHTTTCCTHHHHHHHHHHHHHHHHHHHHHHTTCCCHHHHHHHHHHHHHHHHHHHHHCC
EEEEBEEBBBEBBEEBBEEBBEEBEEEEEEBEEBEEBBEEBBEEBEEBBEEEEEEEEBEEBBEEBEEBBEEBBEBBEEEEEBBEEBEEBBEEBBEBBEEBBEE
MYGKLNHIINHLKDLVNQLNKNWHNDRKNLHDLDQNIKQMFENLDNFMQGKENDGKLQKFMKDLKEMFHQLHNHLQQRRDTVHHFKQEVHHFIHQFKHLVHR
                                                   53     59
```

(a)

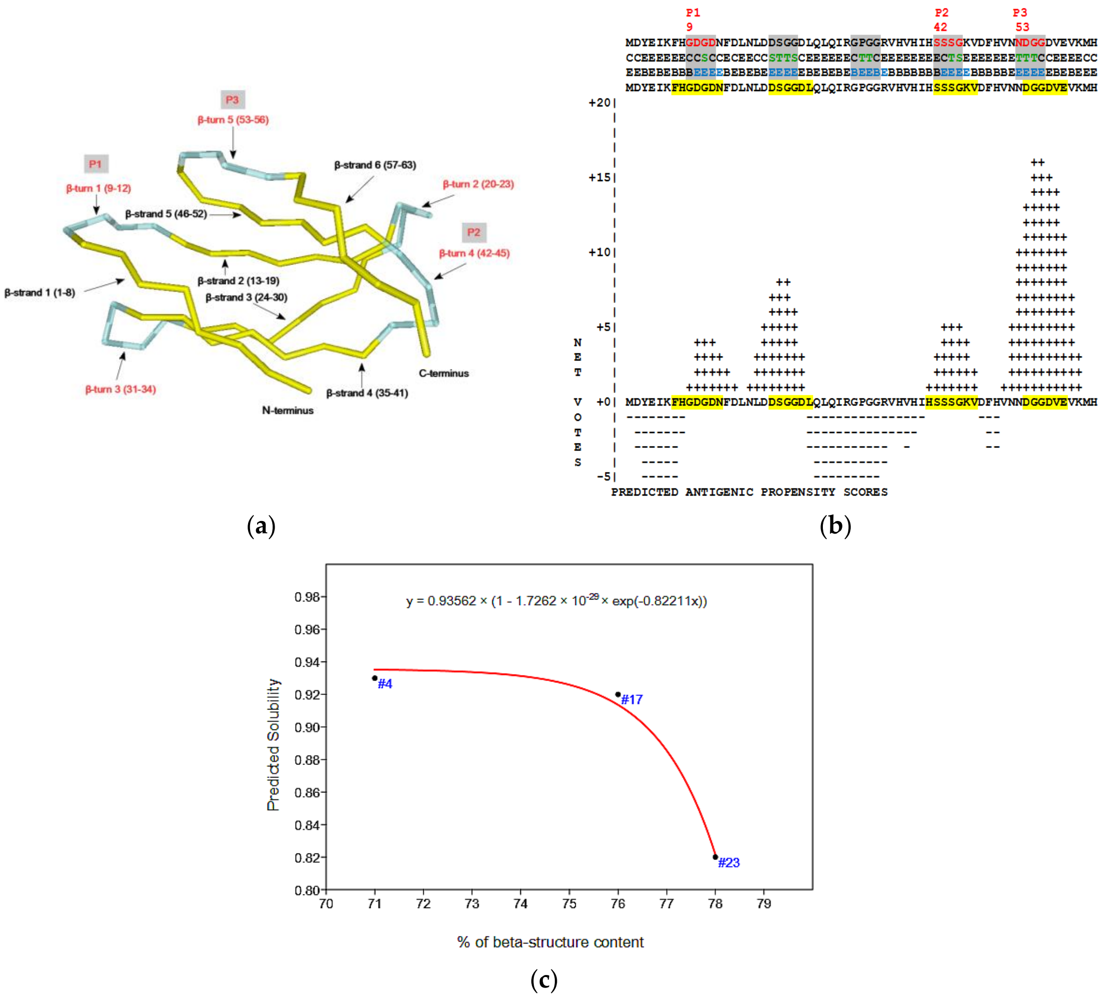
(b)

```
                             SVM LBtop     % Probabilty of
                             Score(C)      correct prediction
49
QGKENDGKLQKFMKD              0.311         60.35
 GKENDGKLQKFMKDL             0.459         65.30
  KENDGKLQKFMKDLK            0.586         69.52
   ENDGKLQKFMKDLKE           0.586         69.52
    NDGKLQKFMKDLKEM          0.382         62.75
     DGKLQKFMKDLKEMF         0.464         65.46
      GKLQKFMKDLKEMFH        0.463         65.43
       KLQKFMKDLKEMFHQ       0.554         68.46
        LQKFMKDLKEMFHQL      0.400         63.33
         QKFMKDLKEMFHQLH     0.596         69.86
          KFMKDLKEMFHQLHN    0.401         63.37
                        73

                             BepiPred      Prediction
                             Score

                        49   0.692  . .    Q|E
                        50   0.855  . .    G|E
                        51   0.979  . .    K|E
                        52   1.362  . .    E|E
                        53   1.304  . .    N|E
                        54   1.261  . .    D|E
                        55   1.081  . .    G|E
                        56   0.593  . .    K|E
```

(c)

Figure 3. (**a**) Surface-exposed regions of Hecht_α rescue protein SynSer3 identified using different methods: Line 1 = Kyte–Doolittle (ExPASy-ProtScale); Line 2 = eight-class secondary structure prediction (SCRATCH-SSpro8, H = α-helix, T = turn, C = the rest); Line 3 = protein surface accessibility (NetSurfP, E = exposed, B = buried); Line 4 = protein antigenicity (SCRATCH-COBEpro); (**b**) 3D SynSer3 structure model using the Phyre2 method, 100% residues modeled at >90% accuracy; (**c**) prediction of SynSer3 linear B-cell epitopes (LBtope, BepiPred).

A typical 3D structure of Hecht_β protein (#17) is presented in Figure 4. The analysis of Hecht_β proteins suggests the existence of three surface exposed regions corresponding to turn 1 (P1), turn 4 (P2) and turn 5 (P3). These positions are predicted to be antigenic regions (Figure 4b). The predicted probability of antigenicity [31] and solubility upon overexpression in *Escherichia coli* [35,36] for all Hecht_α and Hecht_β proteins is presented in Table A3. Those methods enable fast and simple virtual screening for desirable properties [31,35,36]. All 17 Hecht_β proteins and 13 of 15 Hecht_α proteins were predicted to be soluble, using the SOLpro and Periscope methods (Table A3) [35,36]. As for Hecht_α SynRescue proteins, the SOLpro method predicted that two of them, that is SynIlvA2 and SynFes2, were insoluble (Table A3). Periscope, a recently-published method for the quantitative prediction of soluble protein expression in the periplasm of *Escherichia coli*, predicted SynIlvA2 and SynFes2 to be soluble. Regarding the solubility-function relationship, there was no significant difference in the bioactivity between the SynIlvA2 and SynIlvA1 (Table A3, data by Fisher at al. [6]). The same is valid for the SynFes2 and SynFes6, since both of them accumulated iron successfully [6]. The given data imply that the solubility prediction should be used with caution.



```
               P1                               P2         P3
               9                                42         53
       MDYEIKFHGDGDNFDLNLDDSGGDLQLQIRGPGGRVHVHIHSSSGKVDFHVNNDGGDVEVKMH
       CCEEEEEECCSCCECEECCSTTSCEEEEEECTTCEEEEEEEECTSEEEEEEETTTCCEEEECC
       EEBEBEBBBEEEEBEBEBEEEEEBEBEBEBEEEBEBBBBBBBEEEEBBBBBEEEEEEBEBEEE
       MDYEIKFHGDGDNFDLNLDDSGGDLQLQIRGPGGRVHVHIHSSSGKVDFHVNNDGGDVEVKMH
  +20|
     |
     |
     |
     |                                                       ++
  +15|                                                       +++
     |                                                       ++++
     |                                                      +++++
     |                                                      +++++
     |                                                      ++++++
  +10|                                                     +++++++
     |                                                     +++++++
     |                     ++                              +++++++
     |                    +++                              ++++++++
     |                    ++++                             ++++++++
   +5|                   +++++                   +++      +++++++++
N    |          +++      +++++                   ++++     +++++++++
E    |          ++++    +++++++                 +++++     ++++++++++
T    |          +++++   +++++++                 ++++++    ++++++++++
     |         +++++++ ++++++++                +++++++   +++++++++++
V  +0| MDYEIKFHGDGDNFDLNLDDSGGDLQLQIRGPGGRVHVHIHSSSGKVDFHVNNDGGDVEVKMH
O    | --------                ----------------       ---
T    |  -------                ---------------         --
E    |  ------                 -----------  -          --
S    |   -----                  ----------
   -5|   -----                  ---------
      PREDICTED ANTIGENIC PROPENSITY SCORES
```

Figure 4. (**a**) 3D structure of Hecht_β protein #17 according to the FOLDpro prediction method (2D7PA template) [52]; (**b**) surface-exposed regions of Hecht_β protein #17 identified using different methods: Line 1 = Kyte–Doolittle (ExPASy-ProtScale); Line 2 = eight-class secondary structure prediction (SCRATCH-SSpro8, E = extended strand, T = turn, S = bend and C = the rest); Line 3 = protein surface accessibility (NetSurfP, E = exposed, B = buried); Line 4 and graph = protein antigenicity (SCRATCH-COBEpro); (**c**) nonlinear drop of predicted solubility for Hecht_β proteins #4, #17 and #23 (SOLpro) with the increase of β-structure content (von Bertalanffy growth function).

The negative correlation between SOLpro predicted solubility and ANTIGENpro predicted antigenicity was significant for Hecht_α SynRescue proteins (r = −0.596, p = 0.019), but it was insignificant for Hecht_β proteins (r = 0.242, p = 0.349). A very similar result was observed for the correlation between Periscope predicted solubility and ANTIGENpro predicted antigenicity, significant for Hecht_α SynRescue proteins (r = −0.795, p = 0.0004) and insignificant for Hecht_β proteins (r = 0.095, p = 0.717). Solubility has an impact on antigenicity, because low soluble or insoluble antigens tend to form aggregates [53]. Large, insoluble aggregates are more immunogenic than small, soluble molecules [53]. Specific modifications of the analyzed molecules may be obtained by amino acid mutations [6,20]. For example, when lysine mutations are introduced at the ends of the Hecht_β protein #45 in order to disfavor fibrillar structure formation and prevent oligomerization [20], the predicted solubility increases slightly, and the spectral peak indicates a small shift (Figure A4).

2.3. Virtual Spectroscopy and 3D Ligand Binding Prediction of Hecht_α (SynRescue) Proteins

The key goal in synthetic biology is to design and produce novel proteins with a specific structure and function [3]. Screening of the third-generation computational libraries for de novo sequences that function in vivo yielded several Hecht_α sequences, termed SynRescue proteins [3,6], that rescue conditionally lethal mutants of *Escherichia coli* (auxotrophs) [3,6,11]. From a practical standpoint, it would be desirable not only to construct and test a large number of structural patterns in proteins, but also to predict the functional characteristics of the newly-designed molecules. To address this problem, we analyzed 15 SynRescue protein sequences using the informational spectrum method (ISM) [40–44]. This is a virtual spectroscopy method for structure-function analysis of proteins based on the amino acid electron-ion interaction potential (EIIP) in the Rydberg energy units [40–44]. According to the underlying theory, the highest peaks found using this type of analysis represent hot spots, i.e., bioactive parts, of the protein molecule [40–44].

Table 2 presents the results of virtual spectroscopy for seven Hecht_α SynRescue proteins that save *Escherichia coli* auxotrophs with disrupted amino acid enzyme pathways for serine (SerB), glutamate/glutamic acid (GltA) and isoleucine (IlvA).

Table 2. Prediction of the bioactive hot spots in the Hecht_α SynSer, SynGlt and SynIlv rescue proteins. LSSA, alternative least-squares spectral analysis.

Synthetic Protein	Spectral Analysis [1] (Fourier Single Series)	Spectral Analysis [1] (LSSA)	Amino Acid/No.	Activity [2] (Cell Growth)
SynSerB3	0.15 and 0.45	0.15 and 0.45	L30 and F92	++
SynSerB1	0.15 and 0.45	0.15 and 0.45	L30 and L92	++
SynSerB4	0.45	0.45	M92	+
SynSerB2	-	-	-	+
SynGltA1	0.16 and 0.34	0.16 and 0.34	M33 and N69	+
SynIlvA1	0.13 and 0.43	0.13 and 0.43	E26 and Q88	+
SynIlvA2	0.13 and 0.43	0.13 and 0.43	E26 and Q88	+

[1] Dominant frequency peak (informational spectrum method (ISM)); - = without a dominant frequency peak; [2] moderate = ++, low = + (Fisher et al. [6]).

2.3.1. SynSerB Rescue Proteins

In their recent study, Digianantonio and Hecht [11] describe the mechanism by which SynSerB3 (Figure 5), a novel regulatory protein discovered in a library of Hecht_α SynRescue sequences, rescues knockout strains of *Escherichia coli*. The newly-constructed protein SynSerB3 provides the necessary function to maintain bacterial cell growth under conditions of *serB* gene deletion, which encodes phosphoserine phosphatase, an enzyme essential for serine biosynthesis [6,11]. The important conclusion made by Digianantonio and Hecht is that de novo proteins, based on the binary coding patterns of the amino acid polarity and a library of *Escherichia coli* sequences, can be used to drive adaptive changes in the gene expression [11]. However, to ensure the rescue function of the artificially-designed SynSerB protein sequences, Hecht and co-workers transformed a large library of

1.5 × 10^6 binary patterned de novo sequences into strains of *Escherichia coli* containing survival gene deletions [6,11].

Our results in Table 2, Figures 5 and 6 show that higher growth rates of auxotrophic/SerB knockout *Escherichia coli* strains [6] exerted by SynSerB3 and SynSerB1 Hecht_α proteins are characterized by two dominant frequency peaks at positions 0.45 (F/L92) and 0.15 (L30), of the periodograms and cross-amplitude (Figure 5a,b and Figure 6). By contrast, lower growth rates of auxotrophic/SerB knockout *Escherichia coli* strains [6] exerted by SynSerB4 and SynSerB2 proteins are characterized by the absence of one or both of the frequency peaks and both cross-amplitude peaks, respectively (Figure 5c,d and Figure 6).

The frequency peaks, i.e., hot spots, of SynSerB3 correspond to the first nonpolar residue of helix 2 (L30) and the central nonpolar residue of the helix 4 (F92) [3]. The same results were obtained by least square spectral analysis (LSSA), as presented in Table 2. Consequently, ISM analysis of de novo protein sequences might be a useful supplementary procedure for the evaluation of potential bioactive sites and selection/virtual screening of novel protein nano-building blocks possessing specific functional characteristics [21].

Additional information related to the possible bioactive sites of the SynRescue protein may be obtained with 3DLigandSite. The method uses predicted de novo protein structures and the ligands present in homologous natural structures to predict ligand binding sites [38,39]. The reactive patterns of the presumed natural ligands (heterogens) may be particularly useful for reconstructing the biochemical pathways of the auxotrophs that the novel proteins could rescue. As an example, the predicted ligand binding sites of the higher activity rescue protein SynSerB3 are given in Figure 7.

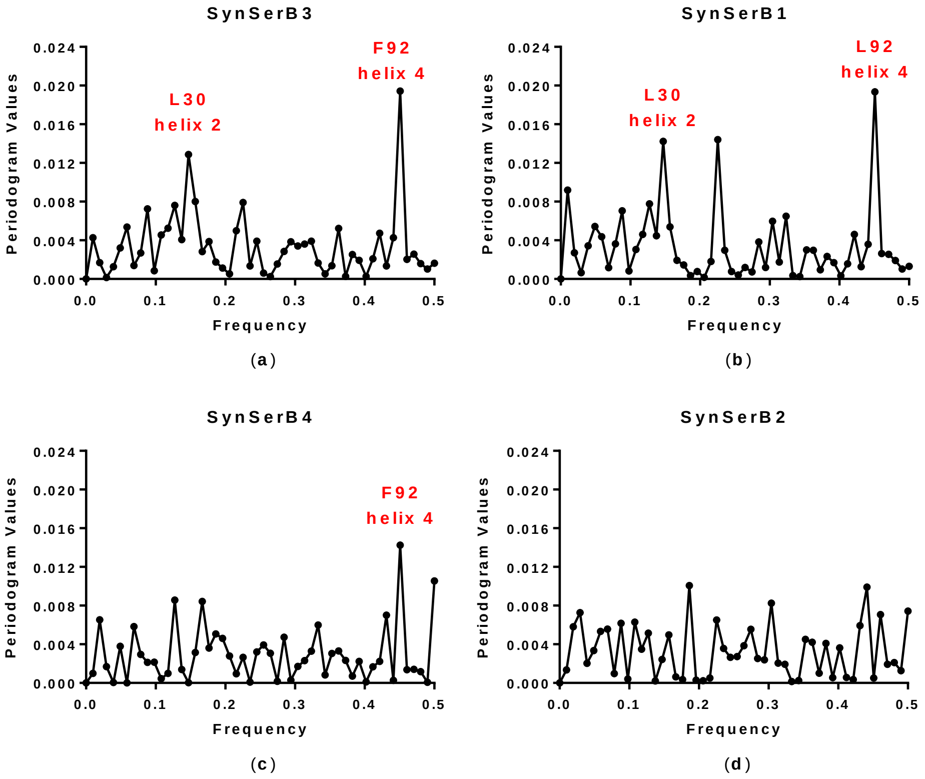


Figure 5. Analysis of de novo SynSerB proteins using the informational spectrum method (ism). Frequency peaks in the periodograms of SynSerB sequences were determined using single series Fourier analysis: (**a**) SynSerB3; (**b**) SynSerB1; (**c**) SynSerB4; and (**d**) SynSerB2.

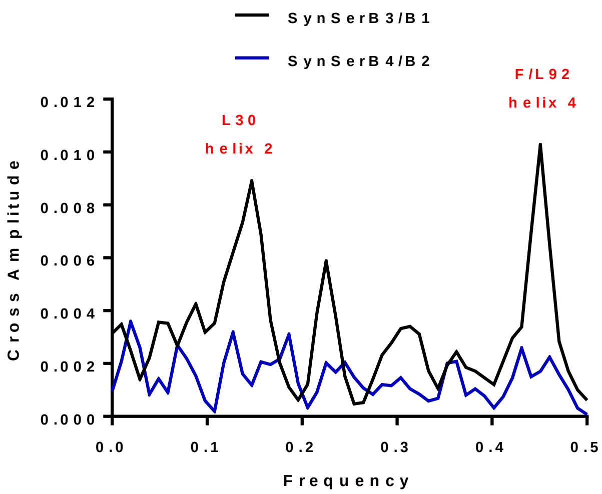
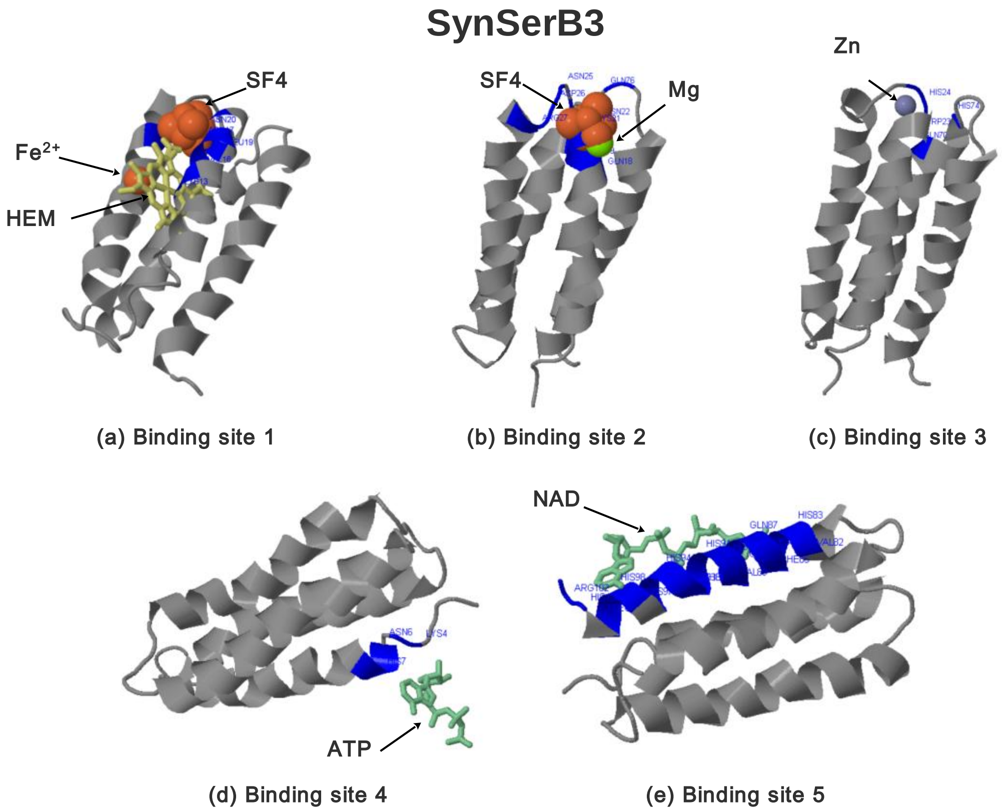


Figure 6. The bivariate Fourier (cross spectrum) analysis of Hecht_α SynSerB1–4 proteins. The electron-ion interaction potential (EIIP) informational spectrum shows two distinct cross-amplitude peaks for the bioactive SynSerB3/SynSerB1rescue proteins at positions 0.15 (L30) and 0.45 (F/L92). Low-activity SynSerB4/SynSerB2 rescue proteins are characterized by the absence of the typical peaks.



(a) Binding site 1

(b) Binding site 2

(c) Binding site 3

(d) Binding site 4

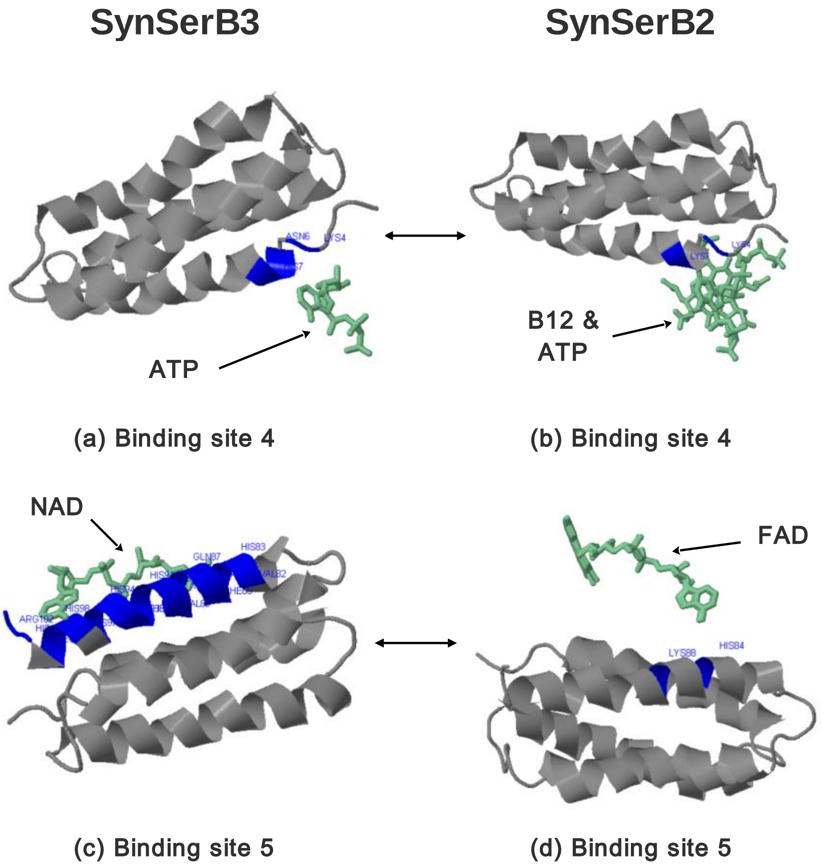
(e) Binding site 5

Figure 7. Heterogens present in the predicted binding sites of SynSerB3 protein using the 3DLigandSite method. (**a**) Binding site 1: heterogens are SF4, HEM (heme) and Fe^{2+}; (**b**) binding site 2: heterogens are SF4 and Mg; (**c**) binding site 3: the heterogen is Zn; (**d**) binding site 4: the heterogen is ATP; (**e**) binding site 5: the heterogen is NAD.

When the binding sites of the higher activity rescue protein SynSerB3 in Figure 7 are compared to the lower activity SynSerB2 mutant, Figure 9 clearly shows that:

- in the SynSerB2 mutant heterogen, FMN binds to binding site 2 instead of SF4/Mg (this region is located between structurally-important stabilizing amino acid positions 26 and 78 [3] of SynRescue proteins, Figure 9a,b);
- FMN interaction with binding site 2 shifts the binding of SF4 to binding site 1, but the additional interaction with HEM (heme) and Fe^{2+} is missing (region 16–31, Figure 9c,d);
- the heterogen B12 in the SynSerB2 mutant disrupts the binding of ATP at binding site 4 (amino acid positions 4–7, Figure 8a,b);
- the heterogen FAD in SynSerB2 mutant disrupts the binding of NAD at binding site 5 (amino acid positions 82–102, Figure 8c,d).

Hecht and co-workers showed that their de novo α-helical proteins frequently exhibit biological functions including the heme binding and peroxidase, esterase and lipase activities [6,10,54,55]. In addition to enzymes, some specific cofactors are involved in metabolic reactions, e.g., adenosine triphosphate (ATP), nicotinamide adenine dinucleotide (NAD), nicotinamide adenine dinucleotide-5-phosphate (NAPD), flavin adenine dinucleotide (FAD), and Fe-protoporphyrin IX (HEM, i.e., heme) [56,57].



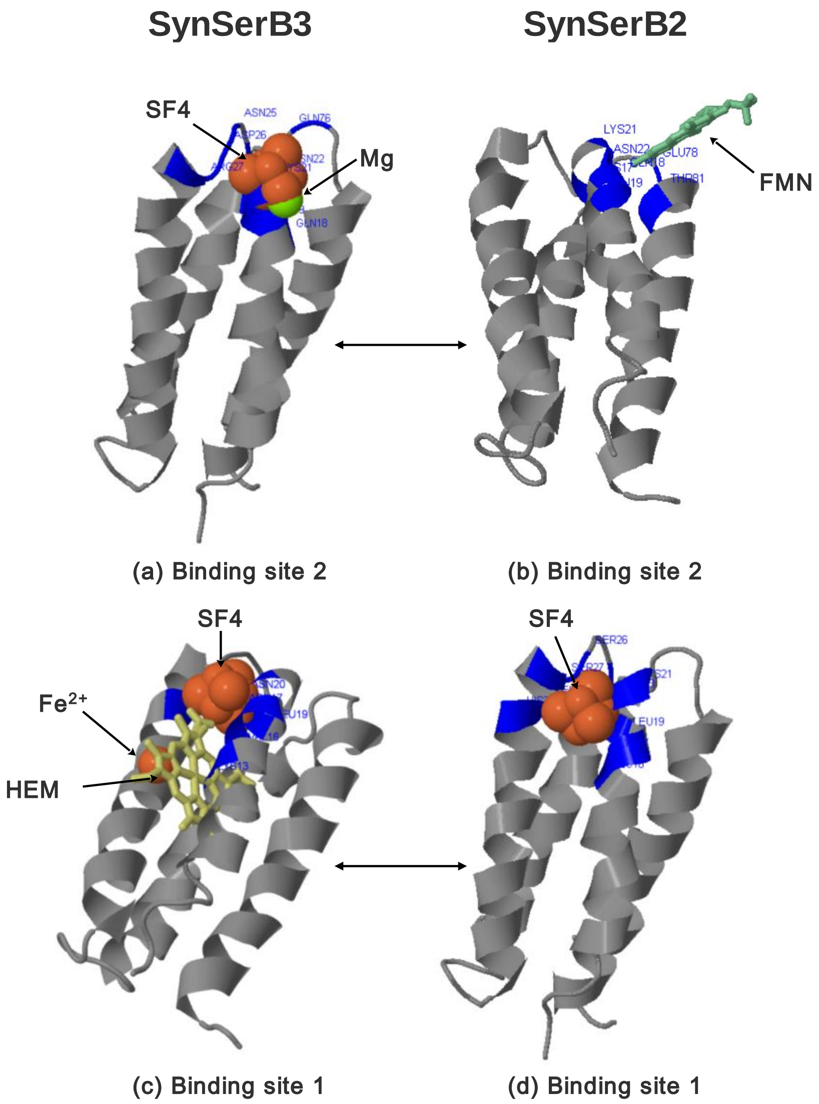
(a) Binding site 4

(b) Binding site 4

(c) Binding site 5

(d) Binding site 5

Figure 8. Heterogens present in the predicted binding sites 4 and 5 of the SynSerB3 and SynSerB2 proteins using the 3DLigandSite method. (**a**) SynSerB3 binding site 4: the heterogen is ATP; (**b**) SynSerB2 binding site 4: heterogens are B12 and ATP; (**c**) SynSerB3 binding site 5: the heterogen is NAD; (**d**) SynSer2 binding site 5: the heterogen is FAD.



(a) Binding site 2 (b) Binding site 2

(c) Binding site 1 (d) Binding site 1

Figure 9. Heterogens present in the predicted binding site 1 of the SynSerB3 and SynSerB2 proteins using the 3DLigandSite method. (**a**) SynSerB3 binding site 2: heterogens are SF4 and Mg; (**b**) SynSerB2 binding site 2: the heterogen is FMN; (**c**) SynSerB3 binding site 1: heterogens are SF4, HEM (heme) and Fe^{2+}; (**d**) SynSerB2 binding site 1: the heterogen is SF4.

The binding assays determined that nearly 66% of the Hecht_α protein sequences of the third generation library bind heme (approximately half at a relatively high level) [10,54]. Of the 80% of the proteins bound that exhibited peroxidase activity, 60% exhibited hydrolase activity and 36% lipase activity [54]. The enzyme activity was up to 10^6-times faster than the uncatalyzed reaction for peroxidase and up to 10^3-times faster than the uncatalyzed reaction for hydrolase and lipase [54]. Hydrolase activities rely on the protein alone, i.e., the enzymatic activity may be found in the absence of a cofactor, as well [10,54]. It is important to note that nearly 30% of heme-binding proteins exhibited some level of enzymatic activity for all functions [54]. In the absence of heme cofactor, esterase and lipase activities were reported in 30% and 20% of the third generation of Hecht_α de novo proteins, respectively, although at lower rates than for natural (evolved) enzymes [10,54].

Using the 3DLigandSite prediction method, the specific binding site was located for several of the cofactors (ATP, NAD, FAD, HEM, SF4). Heme binding is relatively easy to achieve with the de novo design since many peptides and proteins have been designed to bind heme [54]. The interaction of the SynSer3 with heme and Fe^{2+} in the aa region 16–31 was predicted by the COBEpro method (the epitope/exposed region NDRKNLH, aa 25–31) and by ISM (L30); Scheme S3 and Table 2, respectively. The list of predicted continuous epitopes for all Hecht_α proteins is shown in the Schemes S1–S15, using the COBEpro method [32]. It remains to be determined whether the binding of a specific heterogen (e.g., heme) to a stabilizing region (e.g., P1) may destabilize and modify the SynSer3 structure and make other sites of the molecule available to other bioactive heterogens. According to Hecht et al., moderately active de novo heme proteins can serve as starting points for the laboratory-based enzyme evolution and the development of molecules with improved activity [55].

As shown in Table 2 and Figures 7–9, different mutations at the specific positions of the SynSerB rescue protein may account for the bioactivity of SynSerB3 and the inactivity of SynSerB2.

Metabolic enzymes and transcription cofactors participate in transcriptional regulation and represent a direct link between cellular metabolism and regulated gene expression [56,57]. They play an important role in the production of the proteins that are necessary for cellular function, metabolism and gene expression. The variety of biological functions exerted by the SynRescue protein group in *Escherichia coli* auxotrophs may derive from the ability of these synthetic proteins to bind different cofactors. The results presented demonstrate that the combining of ISM and 3DLigandSite methods might be a useful filter for the virtual selection of molecules with desirable properties.

IS-based phylogenetic analysis (ISTREE) [58] of the SynSerBRescue protein clustering and standard phylogenetic analysis of the protein sequences [59,60] provided the same information regarding the directed in vitro evolution of *Escherichia coli* synthetic rescue proteins for serine, i.e., SynSerBRescue (Figure A5). The evolution of synthetic rescue proteins is visualized in Figure A5, from the least active member SynSerB2 (closely related to the parent WA20 structure) to the most active members SynSerB3 and SynSerB1, distant from WA20.

This result confirms that at the level of phylogenetic/molecular evolution analysis, two different analytical methods, amino acid electron-ion interaction potential (ISTREE) and amino acid molecular homologies (TreeDyn), render identical conclusions.

2.3.2. SynIlvA Rescue Proteins

Another important example of further synthetic biology investigation is the auxotrophic/IlvA knockout *Escherichia coli* strains that have disrupted encoding of threonine deaminase. This enzyme catalyzes the first step in the production of isoleucine from threonine. Fisher et al. [6] report that functional SynIlvA1 rescue proteins lose the ability to save auxotrophs upon K to A mutation at the amino acid position 42. The A42 mutants of SynIlvA1 lose the ability to rescue *Escherichia coli* auxotrophs despite the fact that there is no clear difference in the information spectrum changes (Figure 10; EIIP values of K and A are almost identical at 0.0371 and 0.0373, respectively). 3D and secondary structures, as modeled by the Phyre2 and 3DLigandSite methods are also identical (Figure 11, Figure A6, respectively). In A42 mutants, the iron/sulfur cluster-SF4 [61] necessary for the deaminase function remains intact at positions E26, S27 and L71, i.e., in the vicinity of the structurally important stabilizing region E26 and K78 (Figure 11b) [3]. However, the 3DLigandSite method shows marked differences in heterogen Fe binding at the amino acid positions 9, 96 and 12, which is present in biologically-active SynIlvA1 (Figure 11c,d) and absent in the inactive A42 mutants of SynIlvA1. This seems to be in line with the reported functional promiscuity of SynIlvA1, "which was originally selected for its ability to rescue the isoleucine auxotroph Δ*ilv* but also rescues the Δ*fes* auxotroph, which is essential for the accumulation of iron" [3]. Therefore, the lack of iron heterogen at specific positions of the molecule could be relevant for the loss of A42 mutant function in SynIlvA1. At this point, it seems reasonable to investigate further the use of the 3DLigandSite method for evaluating the impact of individual mutations on protein function and the directed evolution of novel SynRescue proteins [55].

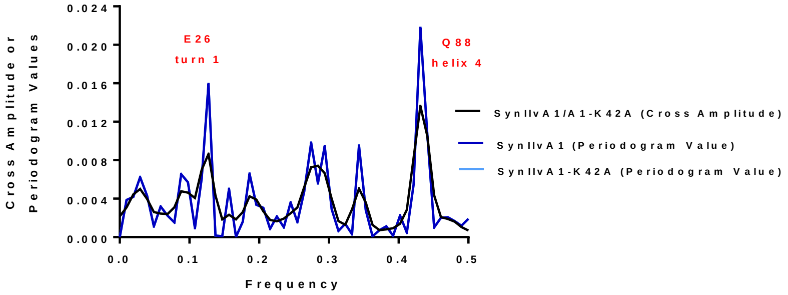
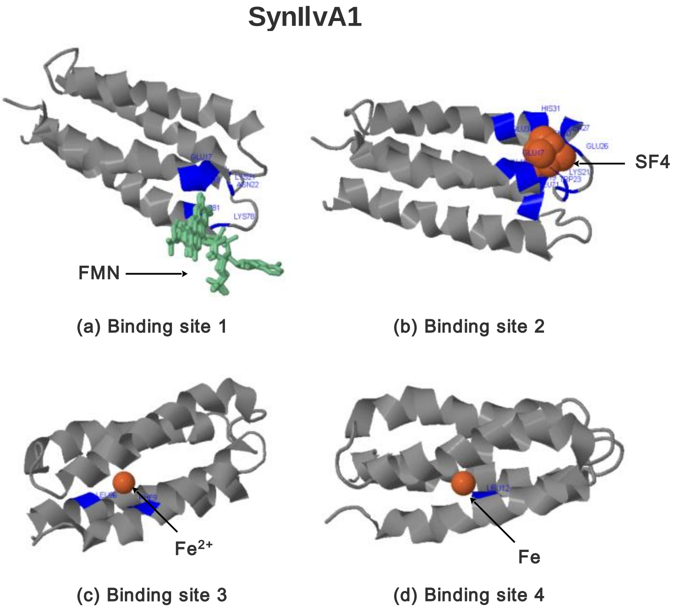


Figure 10. Analysis of Hecht_α SynIlvA1 and SynIlvA1-K42A rescue proteins using the informational spectrum method (ISM) based on EIIP. Cross-amplitude values and individual periodogram values have identical frequency peaks at positions 0.13 (E26) and 0.43 (Q88).



(a) Binding site 1

(b) Binding site 2

(c) Binding site 3

(d) Binding site 4

Figure 11. Heterogens present in the predicted binding sites of SynIlvA1 rescue protein using the 3DLigandSite method. (**a**) Binding site 1: the heterogen is FMN; (**b**) binding site 2: the heterogen is SF4; (**c**) binding site 3: the heterogen is Fe^{2+}; (**d**) binding site 4: the heterogen is Fe. The K42 → A42 mutant does not have binding sites 3 and 4.

2.3.3. SynFes and SynGltA Rescue Proteins

Fes gene is not involved in biosynthetic pathways. It functions in iron acquisition by encoding enterobactin esterase, which cleaves the iron-bound enterobactin siderophore [6]. This allows cells to acquire iron in iron-limited environments [6]. Fisher et al. report that cells expressing the SynFes6 and SynFes2 rescue proteins accumulate six- and 10-fold more iron than control cells, respectively [6]. The difference in SynFes6 and SynFes2 iron accumulation could be ascribed to three facts detected by the 3DLigandSite method:

- in addition to the positions 13 (Fe/Fe^{2+}) and 49 (Fe^{2+}) shared by SynFes6 and SynFes2, SynFes2 has two extra Fe heterogen binding positions at amino acid sites 64 and 96;
- at position Q49 in SynFes6, FAD and B12 could disrupt the binding of other heterogens (Fe^{2+}, ATP, NAD, GAL, MAN and GLC), which is not the case for SynFes2;
- SynFes2 has two additional binding sites for heterogen Ca at positions 1 and 49.

Table 3 shows the results of analysis for SynFesRescue proteins using EIIP information spectrum (ISM) analysis. The SynFes2 and SynFes6 rescue proteins that accumulate iron show two distinct frequency peaks at positions 26 (turn 1) and 37 (helix 2), Table 3 and Figure 12a. Position 26 belongs to the structurally important stabilizing part of SynRescue proteins [3]. SynFes1 protein does not have the second peak at position 37, i.e., in the molecular EIIP spectrum; the peak of a nonpolar residue close to the central part of helix 2 is missing.

Table 3. Prediction of bioactive hot spots in Hecht_α SynFes rescue proteins.

SynFesRescue	Spectral Analysis [1]	Amino Acid/No.
SynFes2	0.13 and 0.18	K26 and L37
SynFes6	0.13 and 0.18	S26 and L37
SynFes1	0.13	N26
SynFes3	0.09	Q18
SynFes5	0.09	Q18
SynFes7	0.29	Q59
SynFes8	0.29	Q59
SynFes4	0.33	M68

[1] Fourier spectral analysis/LSSA (ISM).

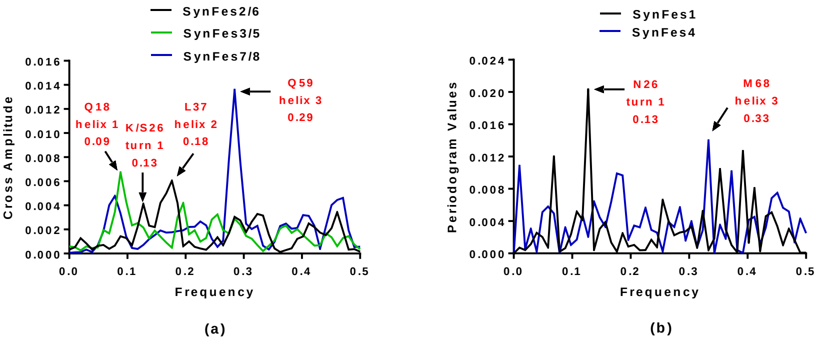


Figure 12. Analysis of SynFes proteins using the informational spectrum method (ISM). Frequency peaks and periodogram values of SynSerB sequences were determined using cross-spectrum (bivariate Fourier) (**a**) and single series Fourier analysis (**b**).

Other members of the SynFesRescue family, SynFes3-5 and SynFes7/8, have one distinct peak at different positions: SynFes3 and SynFes5 at the helix 1 position and SynFes4 and SynFes7/8 at the helix 3 position (Table 3, Figure 12). In a similar way to SynSerBRescue (Figure A5), ISM-based phylogenetic analysis of the SynFesRescue clustering is comparable to standard phylogenetic/molecular evolution in the results for homologous sequences (Figure A7). This suggests that that a combined application of the 3DLigandSite and ISM methods is a useful step in the characterization of synthetic proteins.

The solubility parameter presented in Table A3 might also influence the structural-functional behavior of the artificial sequences, e.g., bioactive and less soluble SynFes2 was reported as forming an extended dimer similar to WA20, which was not the case for the more soluble SynGltA1 that forms a very weakly-associating dimer or an extended monomer [3].

Hecht et al. [62] have recently shown that SynGltA protein acts as a rescuer of *Escherichia coli* cells deleted for *gltA* gene. Deletion of this gene disables the citric acid cycle, and the rescue protein SynGltA restores it [62]. ISM virtual spectroscopy, based on electron-ion interaction potential, predicted multiple bioactive sites located in all four helices of the SynGltA rescue protein (Figure 13). Another method, 3DLigandSite prediction in Figure 14, locates binding sites in SynGltA for several metabolic cofactors (ATP, NAD, FAD) that are of importance for the citric acid cycle [57]. These findings are in line with the observation of Hecht et al. [62] that non-natural rescue proteins recover energy metabolism by activating alternative metabolic pathways.

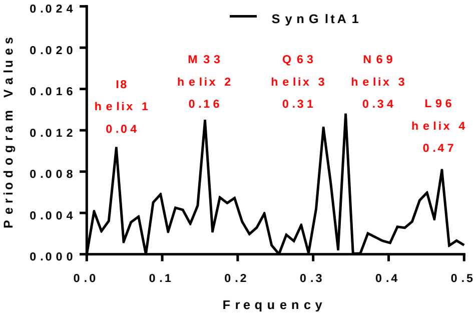
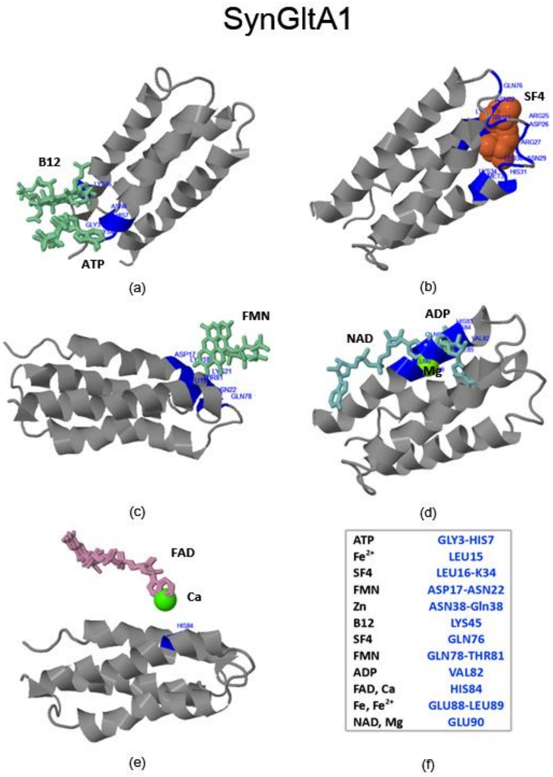


Figure 13. Analysis of rescue SynGltA1 protein using the informational spectrum method (ISM). Frequency peaks in the periodograms of the SynGltA1 sequence were determined using single series Fourier analysis.



ATP	GLY3-HIS7
Fe^{2+}	LEU15
SF4	LEU16-K34
FMN	ASP17-ASN22
Zn	ASN38-Gln38
B12	LYS45
SF4	GLN76
FMN	GLN78-THR81
ADP	VAL82
FAD, Ca	HIS84
Fe, Fe^{2+}	GLU88-LEU89
NAD, Mg	GLU90

Figure 14. Heterogens present in the predicted binding sites of SynGltA1 protein using the 3DLigandSite method. (**a**) Binding site 1: heterogens are B12 and ATP; (**b**) binding site 2: the heterogen is SF4; (**c**) binding site 3: the heterogen is FMN; (**d**) binding site 4: heterogens are NAD, ADP and Mg; (**e**) binding site 5: heterogens are FAD and Ca.

2.4. Virtual Spectroscopy and 3D Structure Prediction of Hecht_β Proteins

Like the Hecht_α dataset, the Hecht_β dataset of de novo protein sequences was analyzed using the informational spectrum method (ISM). This virtual spectroscopy method is useful for structure/function analysis of proteins and the identification of functional protein domains [40–44]. The method is also applicable for the assessment of biological effects of mutations. Frequency peaks of the EIIP periodograms denote the important parts of the molecules. ISM analysis of the Hecht_β dataset, presented in Figure 15, shows that the sequences cluster into five different subgroups, each of them having a distinct peak at a different part of the β-protein. Those peaks identify important regions that predict bioactive epitopes (Figures 2b and 4a,b, Table A4 and Schemes S16–S32).

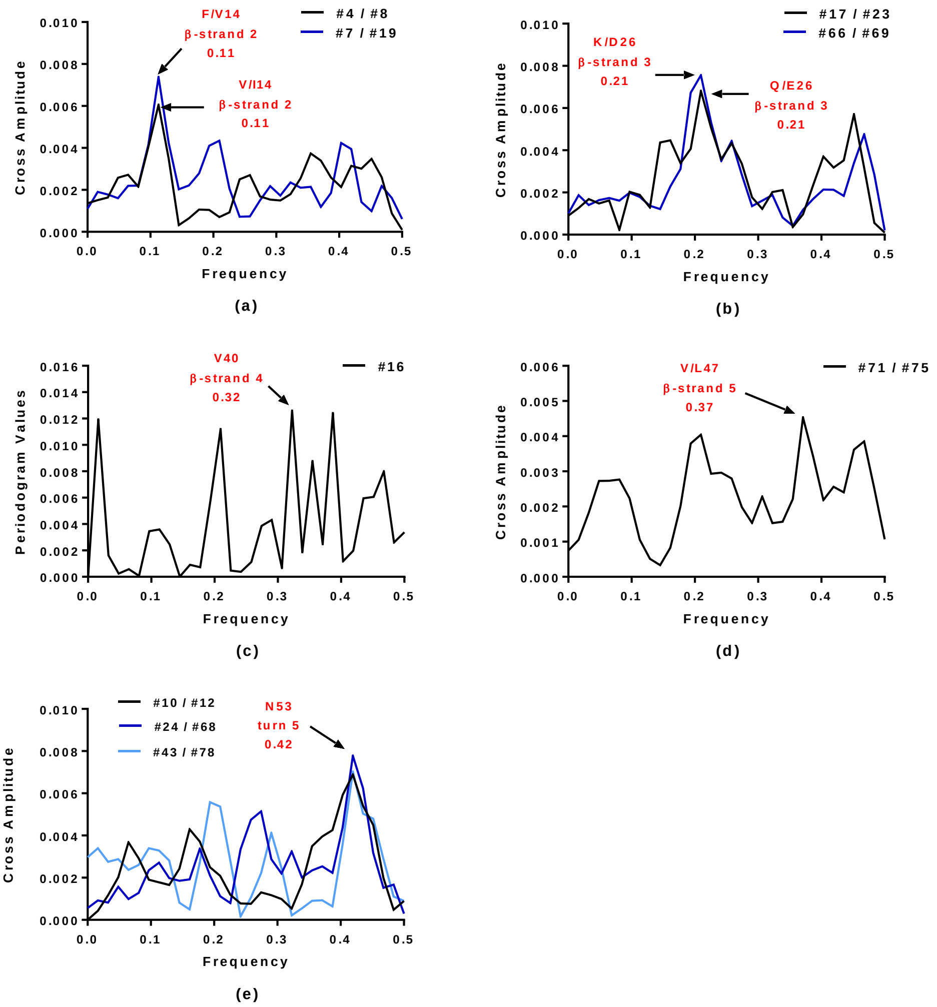


Figure 15. Analysis of Hecht_β proteins with the informational spectrum method (ISM). Frequency peaks of the periodograms denote important functional parts of the molecules related to continuous epitopes. (**a**) Cross amplitudes of Hecht_β proteins: #4 and #8, #7 and #19; (**b**) Cross amplitudes of Hecht_β proteins: #17 and #23, #66 and #69; (**c**) Characteristic peaks of Hecht_β protein #16; (**d**) Cross amplitude of Hecht_β proteins #71 and #75; (**e**) Cross amplitudes of Hecht_β proteins: #10 and #12, #24 and #68, #43 and #78.

The hydropathy of Hecht_α and Hecht_β proteins was also investigated using the sliding block based on the Kyte–Doolittle scale. Figure 2b shows that Hecht_β proteins are characterized by N-terminus and C-terminus epitopes and that the central part in the vicinity of turn 3 is buried (Figures 2b and 4a,b). ISM analysis in Figure 15 complements the Kyte–Doolittle method for the epitope detection (e.g., for N-terminus P1 detection in Figure 2b) and offers two simple rules for the antigenic site location, as follows:

If there are two N-terminus epitopes, then the peak 1 (0.11/aa14) and the peak 2 (0.21/aa26) are in the vicinity of the epitope 1 and epitope 2 ends, respectively. If Hecht_β protein has only one N-terminus epitope, peak 1 (0.11/aa14) is the central part of the antigenic site, and peak 2 (0.21/aa26) is at the epitope end.

At the C-terminal end, there are also two possible epitopes. The peak 0.42 (aa53) is located within epitope 1, situated at the very end of the protein sequence (Figure 15, Table A4 and Schemes S16–32). It corresponds to the antigenic region P3 (aa53–56) determined by the Kyte–Doolittle method (Figure 2b). The peaks 0.32 (aa40) and 0.37 (aa47) are within C-terminus epitope 2 and correspond to the region P2 (aa42–45) predicted by the Kyte–Doolittle method (Figure 2b).

A typical example for the epitope location is protein #17 presented in Figure 4b. The results of ISM spectral analysis of Hecht_β proteins were tested by COBEpro detection of continuous epitopes given in Schemes S16–S32 and Table A4. The results are consistent with the fact that ISM virtual spectroscopy detects bioactive protein regions [40–44].

The 3D structure model of a typical Hecht_β protein #17 is presented in Figure 4a. The FOLDpro method was used for β-protein fold recognition and template-based 3D structure prediction, since the Phyre2 server could not predict the 3D structure and ligand binding with sufficient precision [37,52]. The FOLDpro method extracted human filamin C template protein 2D7PA in 12 Hecht_β proteins (#4, #7, #8, #16, #17,#19, #23, #24, #66, #71, #75, #78; Scheme S33) and human filamin C protein 2D7NA in four Hecht_β proteins (#12, #43, #68, #69). For protein #10, the template is the 2A4CA protein (mouse cadherin-11).

2.5. Structural and Functional Characterization of Hecht_α and Hecht_β Proteins

As discussed by Woolfson et al. [2,5], de novo protein design is closely related to the synthetic biology approach to producing standard sets of polypeptide components, which are designed to solve problems across different biological systems. If properly standardized, those components can be applied in a modular manner to different biochemical problems [5].

The results of virtual spectroscopy, hydropathy analysis and structure-function modeling based on the Hecht_α and Hecht_β protein dataset imply that the proposed methods could be used for the virtual screening of artificial proteins. Additional investigations on different and varied datasets are needed to confirm the general applicability of this concept. The structure elucidation of proteins using NMR and crystallography is a slow and expensive process. It is estimated that the cost of determining each new structure is in the order of $100,000 [63]. The number of known protein sequences is about 400-times larger than the number of experimentally-determined structures, and the number of new sequences grows much faster than the number of structures [64]. However, the cost of computer modeling is much lower (on average $10 per compound [63]), which explains why the computational methods for protein structure and function prediction are important.

Our analysis of the structural-functional relationships and directed evolution of Hecht_α and Hecht_β proteins in *Escherichia coli* is in line with the new approach of Petoukhov [65] “for modeling genetically specified structures and processes in living organisms using mathematical tools of the theory of resonances”. The analysis of the physico-chemical properties of amino acids related to the codon information values and transition-probability distributions in short-term evolution, as discussed by Jiménez-Montano et al. [66], could additionally contribute to a better understanding of how de novo-designed proteins can drive adaptive changes in gene expression in order to provide life-sustaining regulatory functions [11].

3. Materials and Methods

3.1. Protein Datasets

The α-protein dataset consisted of 15 de novo artificial proteins constructed by Hecht et al. (Hecht-α) [6], using a combinatorial library of *Escherichia coli* sequences designed to fold into 102-residue 4-helix bundles (Table S1). The synthetic genes were made using degenerate DNA codons [3,6]:

- VAN (V = A, C, G) was used to encode six polar residues (H, Q, N, K, D, E) and
- NTN (N = A, T, C, G) was used to encode five nonpolar residues (F, L, I, M, V).

Neutral amino acids, with the exception of alanine (A) and cysteine (C), were occasionally used, according to the specificity of the helix/turn protein design [9]. The amino acid septapeptide pattern pnpppnnp, consisting of polar (p) and nonpolar (n) residues, served to approximate an α-structural repeat of the 3.6 residue/turn [9]. The list of α-sequences is given in Table S1.

The β-protein dataset consisted of 17 de novo 6-stranded β-sheet proteins designed by Hecht and coworkers (Hecht_ β) [9,12] using a combinatorial library of synthetic genes. The list of β-sequences is given in Table S2. As with the α-sequences, the combinatorial diversity of the 63-residue-long synthetic β-proteins was achieved by allowing degenerate codons to specify 6 polar (H, Q, N, K, D, E) and 4 nonpolar residues (F, L, I, V) [12]. Each β-protein consisted of 6 polar(p) and nonpolar(n) amino acid septapeptide patterns pnpnpnp [9,12]. The 6 strands were punctuated by 5 turns, made up of 4 amino acid residues each [12].

3.2. Spectral Analysis

The periodicity in de novo α-helical and β-sheet protein structures presented in Table 1 was investigated using the normalized PRIFT method of Cornette et al. [27], which is based on the results of 38 published hydrophobicity scales compared for their ability to identify the characteristic periods of helices/turns (Table A2). The informational spectrum method (ISM) based on electron-ion interaction potential (EIIP) was used to analyze the bioactivity of de novo α- and β-proteins (Tables 2 and 3) [40–44]. The values of EIIP for 20 amino acids are given in Table A2.

Primary amino acid sequences of 15 de novo α-proteins and 17 de novo β-proteins, presented in Tables S1 and S2, were converted into a numerical series by assigning the normalized PRIFT and EIIP value to each amino acid [27,40–44]. The PRIFT and EIIP spectrum (ISM) of each protein sequence was calculated by means of a Fourier spectral analysis and least-squares spectral analysis (LSSA) in order to obtain the highest frequency peaks of the periodogram [27,40–44]. These peaks, i.e., hot spots, denote the structural or bioactive part of the molecule, according to the theoretical concepts of PRIFT and ISM, respectively [27,40–44]. Peak position = 2 × frequency × sequence length. Software STATISTICA for Windows Version 7.0 was used for the Fourier analysis [67] and PAST software Version 3.14 for least-squares analysis [68].

Least-squares spectral analysis (LSSA) estimates a frequency spectrum using a least squares fit of sinusoids to data samples [68–70]. The method gives similar results as Fourier spectral analysis, but is more resistant to noise and appropriate if the time series is long enough to contain at least four cycles [68]. The frequency axis is in units of 1/(x unit). The power axis is in units proportional to the square of the amplitudes of the sinusoids present in the data [68,69].

3.3. Bioinformatic Software Tools Used for Sequence Analyzes

3.3.1. Hydrophobicity Profiles

Surface-exposed regions of de novo α- and β-proteins presented in Figure 2 were identified using the Kyte–Doolittle scale (Table A2) [28,50]. The analyses were based on 3-point moving average values of the 9 amino acid sliding blocks (Figure 2) [28,67,68]. The ExPASy-ProtScale software tool of the ExPASy SIB Bioinformatics Resource Portal was used to compute and represent the amino acid profiles produced by the scale [28].

3.3.2. Solubility, Antigenicity, Surface Accessibility, 2D/3D and Tree Structure Predictions

- The protein 2D structure prediction in Figure 3a was carried out using the SSpro8 server, which adopts the full DSSP 8-class output classification [29]: H = α-helix, G = 3–10-helix, I = pi-helix (extremely rare), E = extended strand, B = β-bridge, T = turn, S = bend and C = the rest.
- The surface/solvent accessibility of amino acids in an amino acid sequence was predicted with the NetSurfP server (E = exposed, B = buried, Figure 3a) [30].
- The probability of protein antigenicity was determined by ANTIGENpro, a sequence-based, alignment-free and pathogen-independent predictor of the protein antigenicity (Table A3) [31]. Prediction of linear B-cell epitopes was carried out using COBEpro, BepiPred and LBotope servers (Figure 3c) [32–34].

- Solubility upon overexpression in *Escherichia coli* was calculated with the SOLpro and Periscope methods (Table A3) [35,36].
- Coupled Phyre2 and 3DLigandSite servers were used to predict 3D structures and protein binding sites, respectively (Figures 3b, 7, 9 and 8, Figures 11 and 14) [37–39]. One hundred percent of Hecht_α protein residues was modeled at >90% confidence.
- The Phyre2 server could not predict the 3D structure of the Hecht_β proteins because the models were insufficiently valid. The confidence was considered too low (<70%) for submission to 3DLigandSite [37]. The FOLDpro method was used for protein fold recognition and template-based 3D structure prediction (Figure 4a, Figure A4) of all β-proteins [52]. The protein 2D and 3D structures were presented using Unipro UGENE software [71]. PDB files of the #17 and #45 models are supplied as Schemes S33–S37.
- The informational spectrum-based phylogenetic analysis in Figures A5a and A7a was done with the ISTREE web service tool [58] and the phylogenetic analysis in Figures A5b and A7b with the Phylogeny.fr platform (TreeDyn) [59,60].

4. Conclusions

De novo proteins designed by Hecht and co-workers [6,9,12] represent structurally and functionally well-characterized subset of α-helical and β-sheet proteins. This dataset may be successfully used both for testing the current methods for the analysis of artificially-designed molecules based on the specific binary patterns of amino acid polarity and for developing the new ones.

The comparative investigations of the bioinformatics methods on the datasets of both de novo and natural proteins may lead to:

1. improvement of the existing tools for protein structure and function analysis,
2. new algorithms for the construction of de novo protein subsets and
3. additional information on the complex natural sequence space and its relation to the individual subspaces of de novo sequences.

Supplementary Materials: The following are available online at http://www.mdpi.com/2078-2489/8/1/29/s1: Table S1: The Hecht_alpha protein dataset consisting of 15 proteins, from Fisher et al. [6]; Table S2: The Hecht_beta protein dataset consisting of 17 proteins, from West et al. [12]; Scheme S1: Predicted epitopes of Hecht_alpha protein SynSerB1 using the COBEpro method [32]; Scheme S2: Predicted epitopes of Hecht_alpha protein SynSerB2 using the COBEpro method [32]; Scheme S3: Predicted epitopes of Hecht_alpha protein SynSerB3 using the COBEpro method [32]; Scheme S4: Predicted epitopes of Hecht_alpha protein SynSerB4 using the COBEpro method [32]; Scheme S5: Predicted epitopes of Hecht_alpha protein SynGltA1 using the COBEpro method [32]; Scheme S6: Predicted epitopes of Hecht_alpha protein SynIlvA1 using the COBEpro method [32]; Scheme S7: Predicted epitopes of Hecht_alpha protein SynIlvA2 using the COBEpro method [32]; Scheme S8: Predicted epitopes of Hecht_alpha protein SynFes1 using the COBEpro method [32]; Scheme S9: Predicted epitopes of Hecht_alpha protein SynFes2 using the COBEpro method [32]; Scheme S10: Predicted epitopes of Hecht_alpha protein SynFes3 using the COBEpro method [32]; Scheme S11: Predicted epitopes of Hecht_alpha protein SynFes4 using the COBEpro method [32]; Scheme S12: Predicted epitopes of Hecht_alpha protein SynFes5 using the COBEpro method [32]; Scheme S13: Predicted epitopes of Hecht_alpha protein SynFes6 using the COBEpro method [32]; Scheme S14: Predicted epitopes of Hecht_alpha protein SynFes7 using the COBEpro method [32]; Scheme S15: Predicted epitopes of Hecht_alpha protein SynFes8 using the COBEpro method [32]; Scheme S16: Predicted epitopes of Hecht_beta protein #4 using the COBEpro method [32]; Scheme S17: Predicted epitopes of Hecht_beta protein #7 using the COBEpro method [32]; Scheme S18: Predicted epitopes of Hecht_beta protein #8 using the COBEpro method [32]; Scheme S19: Predicted epitopes of Hecht_beta protein #10 using the COBEpro method [32]; Scheme S20: Predicted epitopes of Hecht_beta protein #12 using the COBEpro method [32]; Scheme S21: Predicted epitopes of Hecht_beta protein #16 using the COBEpro method [32]; Scheme S22: Predicted epitopes of Hecht_beta protein #17 using the COBEpro method [32]; Scheme S23: Predicted epitopes of Hecht_beta protein #19 using the COBEpro method [32]; Scheme S24: Predicted epitopes of Hecht_beta protein #23 using the COBEpro method [32]; Scheme S25: Predicted epitopes of Hecht_beta protein #24 using the COBEpro method [32]; Scheme S26: Predicted epitopes of Hecht_beta protein #43 using the COBEpro method [32]; Scheme S27: Predicted epitopes of Hecht_beta protein #66 using the COBEpro method [32]; Scheme S28: Predicted epitopes of Hecht_beta protein #68 using the COBEpro method [32]; Scheme S29: Predicted epitopes of Hecht_beta protein #69 using the COBEpro method [32]; Scheme S30: Predicted epitopes of Hecht_beta protein #71 using the COBEpro method [32]; Scheme S31: Predicted epitopes of Hecht_beta protein #75 using the COBEpro method [32]; Scheme S32: Predicted epitopes of Hecht_beta protein #78 using the COBEpro method [32]; Scheme S33:

Hecht_beta #17.pdb; Scheme S34: Hecht_beta #45.pdb; Scheme S35: Hecht_beta #45mutF82.pdb; Scheme S36: Hecht_beta #45mutV5.pdb; Scheme S37: Hecht_beta #45mutV5&F82.pdb.

Acknowledgments: The support of the Croatian Ministry of Science, Education and Sports is gratefully acknowledged (Grant No. 098-0982929-2524).

Author Contributions: Nikola Štambuk and Paško Konjevoda analyzed the data, designed and performed the research and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix

Table A1. The genetic code. Twenty amino acids and three stop codons for the protein synthesis are specified by 64 nucleotide triplets. Polar (p) and nonpolar (n) amino acid groups for the de novo design of Hecht-α and Hecht-β protein structures are shown in red and blue, respectively [9,72].

First (5') Letter	**Second Letter**				**Third (3') Letter**
	U/T	**C**	**A**	**G**	
U/T	Phe (F)	Ser (S)	Tyr (Y)	Cys (C)	U/T
	Phe (F)	Ser (S)	Tyr (Y)	Cys (C)	C
	Leu (L)	Ser (S)	stop	stop	A
	Leu (L)	Ser (S)	stop	Trp (W)	G
C	Leu (L)	Pro (P)	His (H)	Arg (R)	U/T
	Leu (L)	Pro (P)	His (H)	Arg (R)	C
	Leu (L)	Pro (P)	Gln (Q)	Arg (R)	A
	Leu (L)	Pro (P)	Gln (Q)	Arg (R)	G
A	Ile (I)	Thr (T)	Asn (N)	Ser (S)	U/T
	Ile (I)	Thr (T)	Asn (N)	Ser (S)	C
	Ile (I)	Thr (T)	Lys (K)	Arg (R)	A
	Met (M)	Thr (T)	Lys (K)	Arg (R)	G
G	Val (V)	Ala (A)	Asp (D)	Gly (G)	U/T
	Val (V)	Ala (A)	Asp (D)	Gly (G)	C
	Val (V)	Ala (A)	Glu (E)	Gly (G)	A
	Val (V)	Ala (A)	Glu (E)	Gly (G)	G

Table A2. The hydrophobicity scales of Cornette et al. [27], Kyte–Doolittle [50] and amino acid electron-ion interaction potential (EIIP) [40–44] used for bioinformatic analyses.

Amino Acid	**Abbreviation**	**Cornette Scale** [1]	**Kyte–Doolittle Scale**	**EIIP (Ry)**
Phenylalanine	Phe (F)	0.140	2.8	0.0946
Leucine	Leu (L)	0.000	3.8	0.0000
Valine	Val (V)	0.114	4.2	0.0057
Isoleucine	Ile (I)	0.102	4.5	0.0000
Methionine	Met (M)	0.164	1.9	0.0823
Serine	Ser (S)	0.699	−0.8	0.0829
Proline	Pro (P)	0.903	−1.6	0.0198
Alanine	Ala (A)	0.622	1.8	0.0373
Threonine	Thr (T)	0.865	−0.7	0.0941
Cysteine	Cys (C)	0.182	2.5	0.0829
Tryptophan	Trp (W)	0.528	−0.9	0.0548
Arginine	Arg (R)	0.485	−4.5	0.0959
Glycine	Gly (G)	0.648	−0.4	0.0050
Tyrosine	Tyr (Y)	0.278	−1.3	0.0516
Histidine	His (H)	0.595	−3.2	0.0242
Glutamine	Gln (Q)	0.970	−3.5	0.0761
Glutamic acid	Glu (E)	0.854	−3.5	0.0058
Asparagine	Asn (N)	0.701	−3.5	0.0036
Aspartic acid	Asp (D)	1.000	−3.5	0.1263
Lysine	Lys (K)	0.995	−3.9	0.0371

[1] Normalized PRIFT.

Table A3. Predicted probability of antigenicity and solubility upon overexpression in *Escherichia coli* for 32 Hecht_α and Hecht_β proteins.

Synthetic Proteins	Predicted Antigenicity [1]	Predicted Solubility [2]	Predicted Solubility [3]
Hecht_α			
SynSerB1	0.65	Soluble (0.94)	Medium (15.28)
SynSerB2	0.65	Soluble (0.78)	Medium (15.34)
SynSerB3	0.56	Soluble (0.93)	Medium (14.75)
SynSerB4	0.45	Soluble (0.97)	Medium (15.30)
SynGltA1	0.52	Soluble (0.92)	Medium (15.17)
SynIlvA1	0.75	Soluble (0.51)	Medium (16.86)
SynIlvA2	0.78	Insoluble (0.51)	Medium (16.38)
SynFes1	0.83	Soluble (0.90)	Medium (15.42)
SynFes2	0.57	Insoluble (0.58)	Medium (15.37)
SynFes3	0.63	Soluble (0.62)	Medium (15.91)
SynFes4	0.81	Soluble (0.50)	Medium (16.85)
SynFes5	0.55	Soluble (0.75)	Medium (15.29)
SynFes6	0.57	Soluble (0.92)	Medium (15.14)
SynFes7	0.42	Soluble (0.98)	Medium (15.40)
SynFes8	0.84	Soluble (0.69)	Medium (16.70)
Hecht_β			
#4	0.84	Soluble (0.93)	Medium (17.76)
#7	0.71	Soluble (0.90)	Medium (20.04)
#8	0.66	Soluble (0.82)	Medium (18.71)
#10	0.58	Soluble (0.90)	Medium (16.49)
#12	0.66	Soluble (0.94)	Medium (16.91)
#16	0.71	Soluble (0.96)	Medium (16.32)
#17	0.79	Soluble (0.92)	Medium (18.24)
#19	0.80	Soluble (0.84)	Medium (17.33)
#23	0.74	Soluble (0.82)	Medium (19.57)
#24	0.66	Soluble (0.91)	Medium (16.49)
#43	0.65	Soluble (0.88)	Medium (20.50)
#66	0.59	Soluble (0.93)	Medium (18.71)
#68	0.32	Soluble (0.94)	Medium (19.18)
#69	0.47	Soluble (0.87)	Medium (18.27)
#71	0.38	Soluble (0.78)	Medium (15.72)
#75	0.51	Soluble (0.80)	Medium (17.58)
#78	0.75	Soluble (0.92)	Medium (21.33)

[1] ANTIGENpro: probability [31]; [2] SOLpro: solubility (probability) [35]; [3] Periscope: expression level value (mg/L) [36].

Table A4. Predicted continuous epitopes of the Hecht_α and Hecht_β proteins.

Hecht_β Protein	Predicted Epitopes N-Terminus [1]	Predicted Epitopes Central [1]	Predicted Epitopes C-Terminus [1]
#4	2	0	1 ‡
#7	2	0	2
#8	1	0	2
#10	2	0	1 †
#12	1	0	1 †
#16	2	0	1 †
#17	2	0	2
#19	2	0	1 †
#23	2	0	2
#24	2	0	1 †
#43	2	0	2
#66	2	0	2
#68	2	0	2
#69	1	0	2
#71	2	0	2
#75	2	0	2
#78	2	0	2

[1] COBEpro [32]; † 1st terminal epitope; ‡ 2nd terminal epitope.

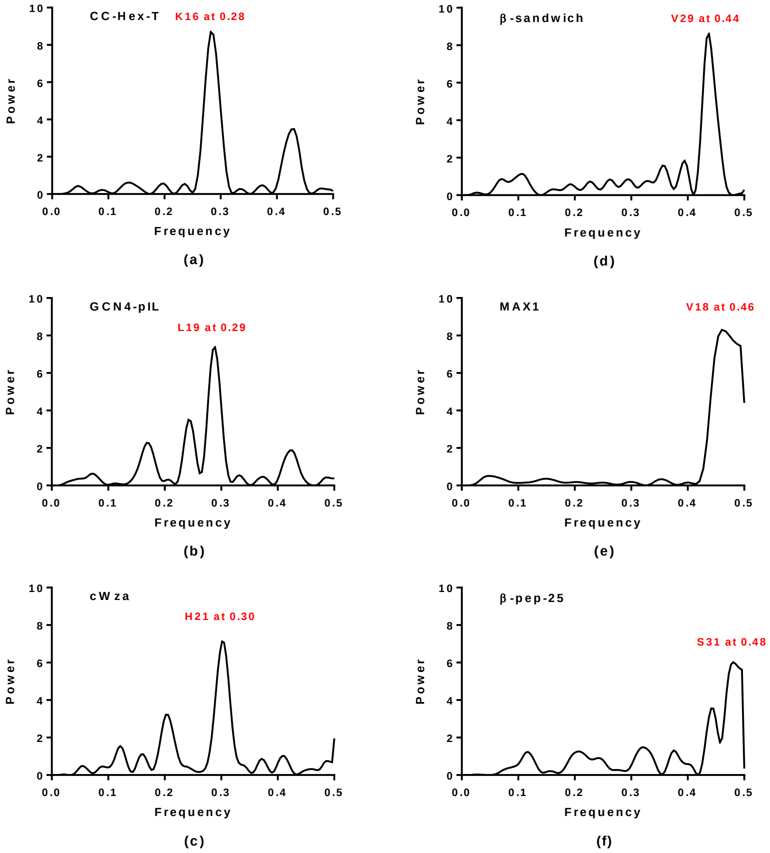


Figure A1. Characteristic peaks of three de novo α-proteins (**a**–**c**) (CC-Hex-T, GCN4-pIL, cWza [23–25]) and three de novo β-proteins (**d**–**f**) (β-sandwich, MAX1, β-pep-25 [47–49]) were determined using the PRIFT/LSSA method of Cornette et al. [27].

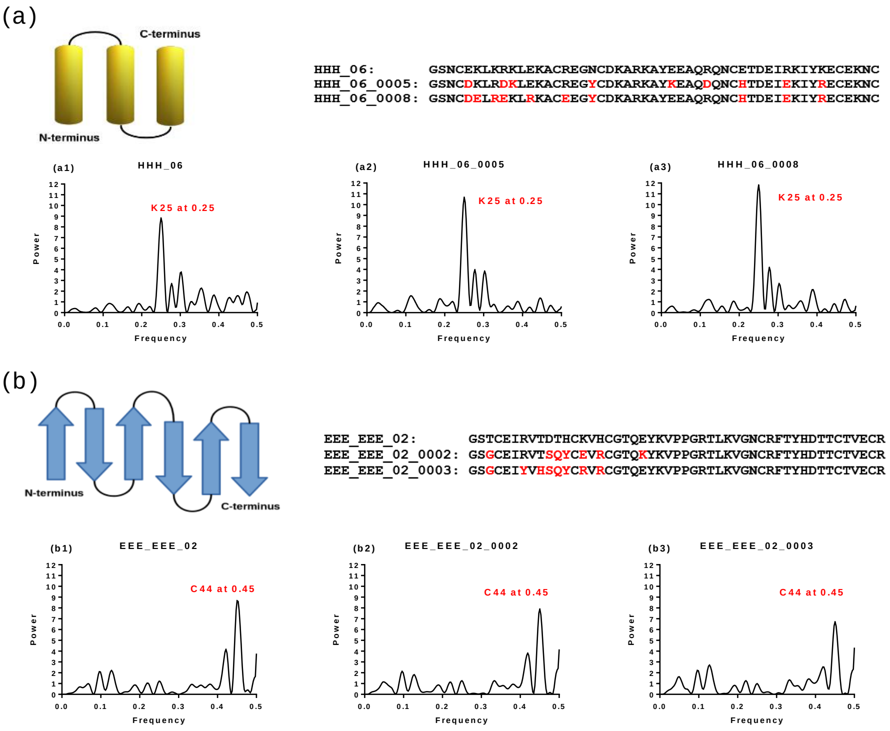


Figure A2. *Cont.*

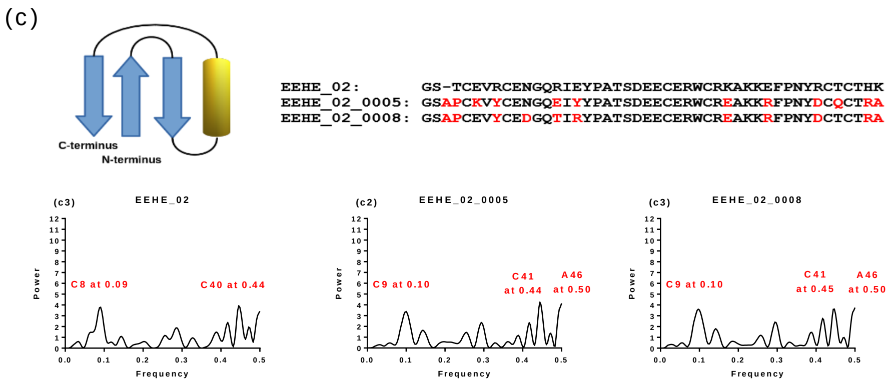
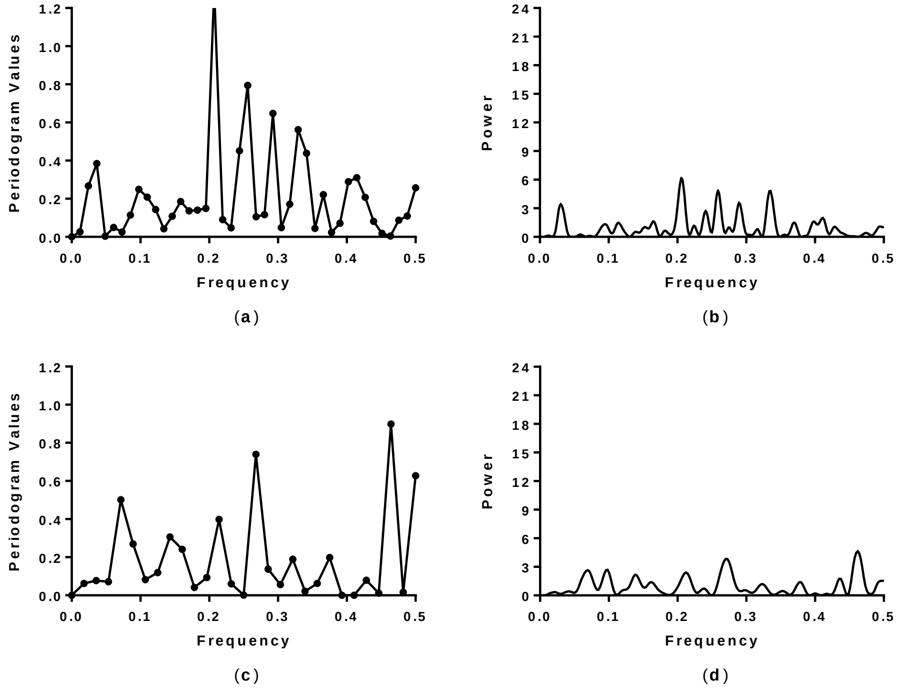


Figure A2. Characteristic frequency peaks of de novo genetically-encodable disulfide-rich peptides and mutants designed by Baker and co-workers [26] were determined using the PRIFT/LSSA method of Cornette et al. [27]. (**a**) α-peptide HHH_06 and its mutants exhibited α-peak at the position 0.25; (**b**) β-peptide EEE_EEE_02 and its mutants exhibited β-peak at the position 0.45; (**c**) mixed class peptide EEHE_02 exhibited two small peaks at positions 0.09 and 0.44. Its mutants EEHE_02_0005 and EEHE_02_0008 were characterized by an additional peak at the position 0.5. Mixed class structures/mutants (**c**) had different distribution of the peaks than when compared to all α-peptides (**a**) and all β-peptides (**b**).



(**a**) (**b**) (**c**) (**d**)

Figure A3. Characteristic peaks of one natural α-protein and one natural β-protein were determined using the PRIFT method of Cornette et al. [27]. (**a**) Natural α-protein 1cc5 did not exhibit the typical α-peak at $x = 0.28$ (Fourier spectral analysis); (**b**) α-protein 1cc5 did not exhibit the typical α-peak at $x = 0.28$ when the alternative method of least-squares spectral analysis was used; (**c**) natural β-protein 1amg-2-AS did not exhibit the typical α-peak at $x = 0.45$ (Fourier spectral analysis); (**d**) natural β-protein 1amg-2-AS did not exhibit the typical α-peak at $x = 0.45$ when the alternative method of least-squares spectral analysis was used.

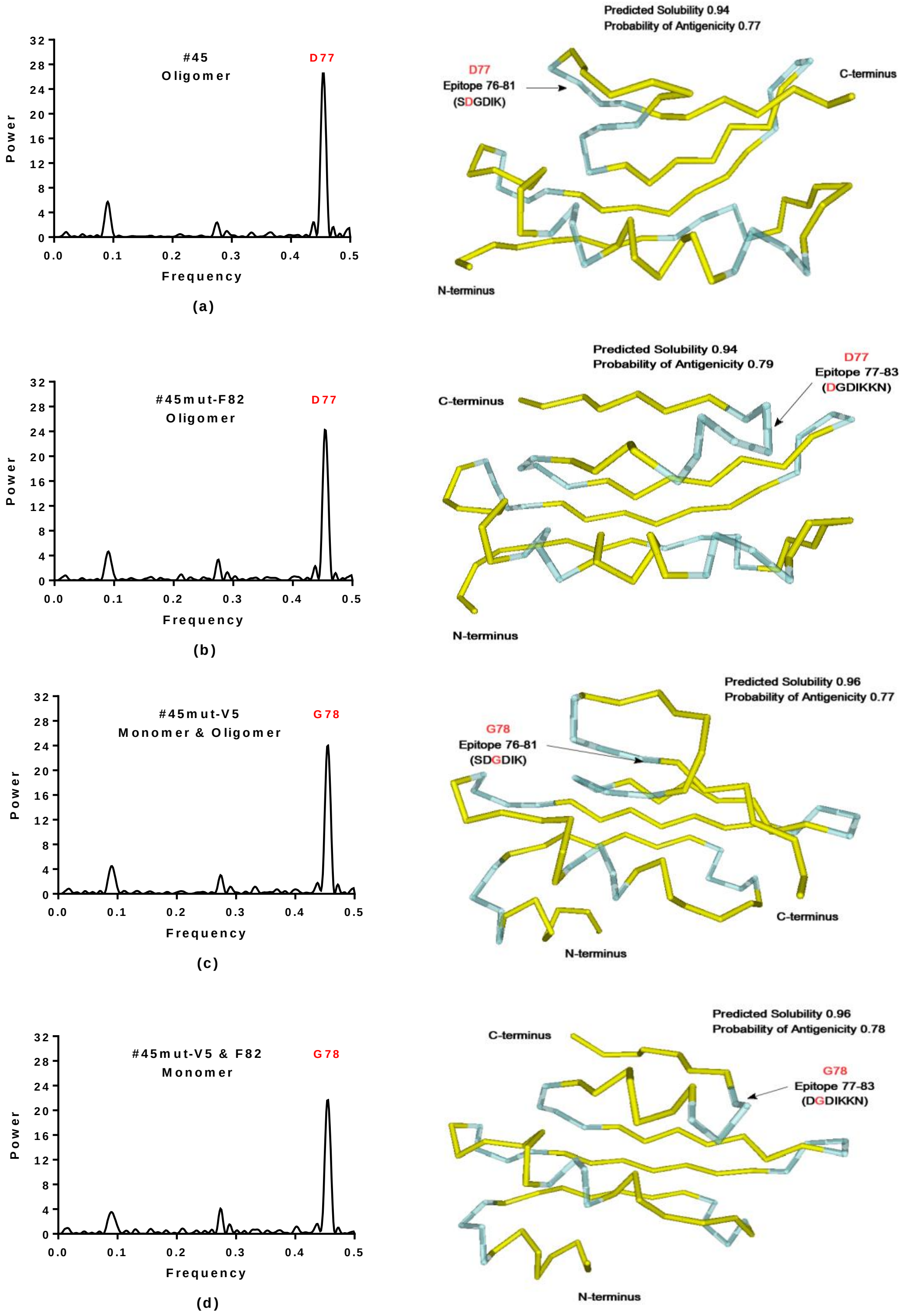
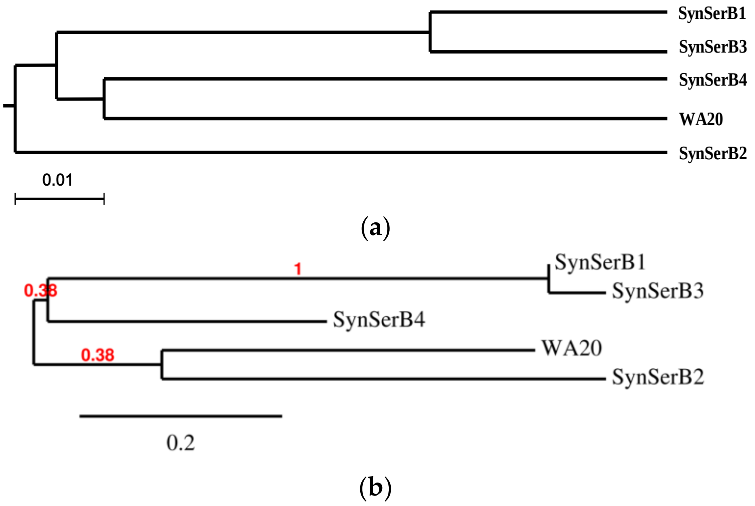


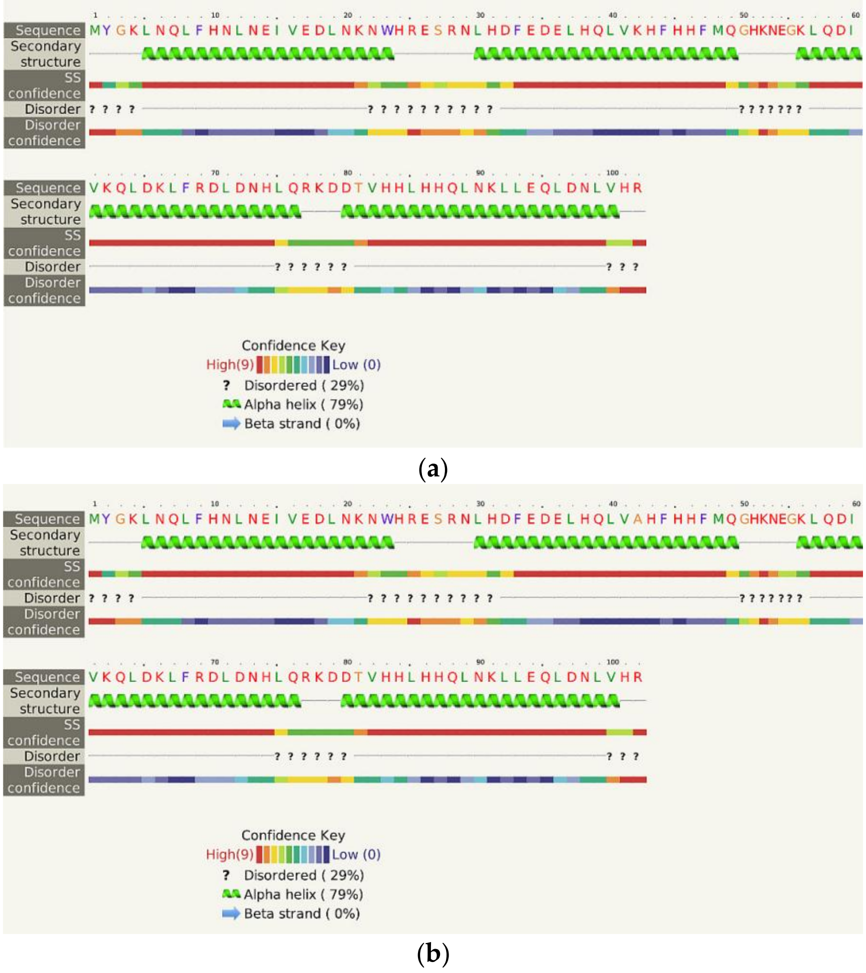
Figure A4. The Spectral analyses and 3D structures of Hecht_β protein #45 and its mutants according to Cornette et al. LSSA [27] and the FOLDpro prediction method (1I58A template, Schemes S34–S37) [52]. Distinct β-peaks at position D77/G78 of turn 7 are located within the most probable epitope predicted using COBEpro. (**a**) #45 oligomer with the distinct β-peak at 0.45/D77; (**b**) #45 mutant F82 → K82, oligomer with the distinct β-peak at 0.45/D77; (**c**) #45 mutant V5 → K5, monomer-oligomer with the distinct β-peak at 0.46/G78; (**d**) #45 mutant V5 and F82 → K5 and K82, monomer with the distinct β-peak at 0.46/G78.



(**a**)

(**b**)

Figure A5. (**a**) Informational spectrum-based phylogenetic analysis of SynSerB1–4 rescue proteins (ISTREE, UPGMA method); (**b**) standard phylogenetic analysis of SynSerB1–4 rescue proteins using the Phylogeny.fr platform.



(**a**)

(**b**)

Figure A6. Identical secondary structures in (**a**) SynIlvA1 Hecht_α protein and (**b**) its A42 mutant, observed by the Phyre2 method.

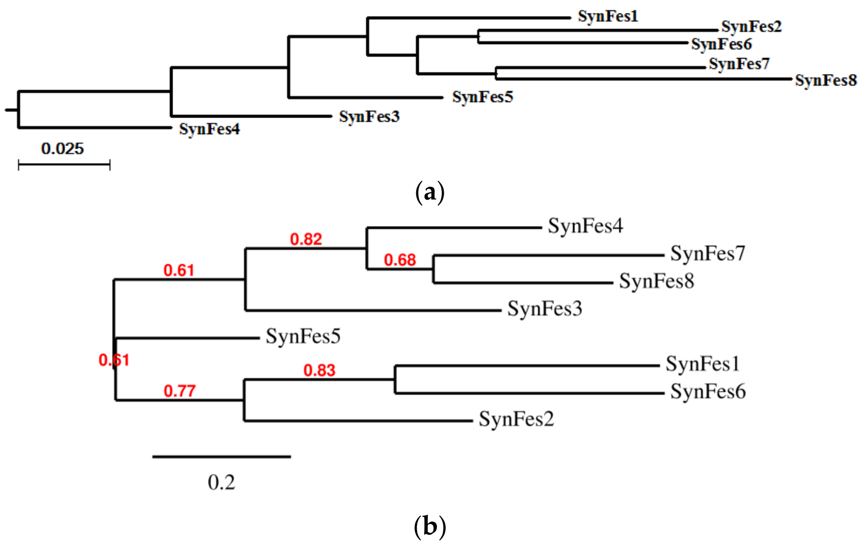


Figure A7. (**a**) Informational spectrum-based phylogenetic analysis of SynFes1–8rescue proteins (ISTREE, NJ method); (**b**) standard phylogenetic analysis of SynFes1–8 proteins using the Phylogeny.fr platform.

References

1. Huang, P.-S.; Boyken, S.E.; Baker, D. The coming of age of de novo protein design. *Nature* **2016**, *537*, 320–327. [CrossRef] [PubMed]
2. Woolfson, D.N.; Bartlett, G.J.; Burton, A.J.; Heal, J.W.; Niitsu, A.; Thomson, A.R.; Wood, C.W. De novo protein design: How do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* **2015**, *33*, 16–26. [CrossRef] [PubMed]
3. Murphy, G.S.; Greisman, J.B.; Hecht, M.H. De novo proteins with life-sustaining functions are structurally dynamic. *J. Mol. Biol.* **2016**, *428*, 399–411. [CrossRef] [PubMed]
4. Woolfson, D.N.; Bartlett, G.J.; Bruning, M.; Thomson, A.R. New currency for old rope: From coiled-coil assemblies to α-helical barrels. *Curr. Opin. Struct. Biol.* **2012**, *22*, 432–441. [CrossRef] [PubMed]
5. Fletcher, J.M.; Boyle, A.I.; Bruning, M.; Bartlett, G.J.; Vincent, T.L.; Zaccai, N.R.; Armstrong, C.T.; Bromley, E.H.C.; Booth, P.J.; Brady, R.L.; et al. A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth. Biol.* **2012**, *1*, 240–250. [CrossRef] [PubMed]
6. Fisher, M.A.; McKinley, K.L.; Bradley, L.H.; Viola, S.R.; Hecht, M.H. De novo designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth. *PLoS ONE* **2011**, *6*, e15364. [CrossRef] [PubMed]
7. Ó Conchúir, S.; Barlow, K.A.; Pache, R.A.; Ollikainen, N.; Kundert, K.; O'Meara, M.J.; Smith, C.A.; Kortemme, T. A web resource for standardized benchmark datasets, metrics, and Rosetta protocols for macromolecular modeling and design. *PLoS ONE* **2015**, *10*, e0130433. [CrossRef] [PubMed]
8. West, M.W.; Hecht, M.H. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.* **1995**, *4*, 2032–2039. [CrossRef] [PubMed]
9. Hecht, M.D.; Das, A.; Go, A.; Bradley, L.H.; Wei, Y. De novo proteins from designed combinatorial libraries. *Protein Sci.* **2004**, *13*, 1711–1723. [CrossRef] [PubMed]
10. Smith, B.A.; Hecht, H. Novel proteins: From fold to function. *Curr. Opin. Chem. Biol.* **2011**, *15*, 421–426. [CrossRef] [PubMed]
11. Digianantonio, K.M.; Hecht, M.H. A protein constructed de novo enables cell growth by altering gene regulation. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 2400–2405. [CrossRef] [PubMed]
12. West, M.W.; Wang, W.; Patterson, J.; Mancias, J.D.; Beasley, J.R.; Hecht, M.H. De novo amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11211–11216. [CrossRef] [PubMed]

13. Štambuk, N.; Konjevoda, P.; Gotovac, N. A new rule-based system for the construction and structural characterization of artificial proteins. In *Chaos and Complex Systems*; Stavrinides, S., Banerjee, S., Caglar, H., Ozer, M., Eds.; Springer: Berlin, Germany, 2013; pp. 95–103.
14. Good, I.G. *Analyzing the Large Number of Variables in Biomedical and Satellite Imagery*, 1st ed.; Wiley: Hoboken, NJ, USA, 2011; pp. 1–3.
15. Bradley, L.H.; Thumfort, P.P.; Hecht, M.H. De novo proteins from binary-patterned combinatorial libraries. In *Protein Design: Methods and Applications*, 1st ed.; Guerois, R., López de la Paz, M., Eds.; Humana Press: Totowa, NJ, USA, 2007; Volume 340, pp. 53–69.
16. Wei, Y.; Liu, T.; Sazinsky, S.L.; Moffet, D.A.; Pelczer, I.; Hecht, M.H. Stably folded de novo proteins from a designed combinatorial library. *Protein Sci.* **2003**, *12*, 92–102. [CrossRef] [PubMed]
17. Moffet, D.A.; Hecht, M.H. De novo proteins from combinatorial libraries. *Chem. Rev.* **2001**, *101*, 3191–3203. [CrossRef] [PubMed]
18. Kamtekar, S.; Schiffer, J.M.; Xiong, H.; Babik, J.M.; Hecht, M.H. Protein design by binary patterning of polar and non-polar amino acids. *Science* **1993**, *262*, 1680–1685. [CrossRef] [PubMed]
19. Rosenbaum, D.M.; Roy, S.; Hecht, M.H. Screening combinatorial libraries of de novo proteins by hydrogen-deuterium exchange and electrospray mass spectrometry. *J. Am. Chem. Soc.* **1999**, *121*, 9509–9513. [CrossRef]
20. Wang, W.; Hecht, M.H. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 2760–2765. [CrossRef] [PubMed]
21. Kobayashi, N.; Yanase, K.; Sato, T.; Unzai, S.; Hecht, M.H.; Arai, R. Self-assembling nano-architectures created from a protein nano-building block using an intermolecularly folded dimeric de novo protein. *J. Am. Chem. Soc.* **2015**, *137*, 11285–11293. [CrossRef] [PubMed]
22. Štambuk, N.; Konjevoda, P. Prediction of secondary protein structure with binary coding patterns of amino acid and nucleotide physicochemical properties. *Int. J. Quant. Chem.* **2003**, *92*, 123–134. [CrossRef]
23. Thomas, F.; Burgess, N.C.; Thomson, A.R.; Woolfson, D.N. Controlling the assembly of coiled–coil peptide nanotubes. *Angew. Chem. Int. Ed.* **2016**, *55*, 987–991. [CrossRef] [PubMed]
24. Burgess, N.C.; Sharp, T.H.; Thomas, F.; Wood, C.W.; Thomson, A.R.; Zaccai, N.R.; Brady, L.; Serpell, L.C.; Woolfson, D.N. Modular design of self-assembling peptide-based nanotubes. *J. Am. Chem. Soc.* **2015**, *137*, 10554–10562. [CrossRef] [PubMed]
25. Mahendran, K.R.; Niitsu, A.; Kong, L.; Thomson, A.R.; Sessions, R.B.; Woolfson, D.N.; Bayley, H. A monodisperse transmembrane α-helical peptide barrel. *Nat. Chem.* **2016**. [CrossRef]
26. Bhardwaj, G.; Mulligan, V.K.; Bahl, D.C.; Gilmore, J.M.; Harvey, P.J.; Cheneval, O.; Buchko, G.; Pulavarti, S.; Kaas, Q.; Eletsky, A.; et al. Accurate de novo design of hyperstable constrained peptides. *Nature* **2016**, *538*, 329–335. [CrossRef] [PubMed]
27. Cornette, J.L.; Cease, K.B.; Margalit, H.; Spouge, J.L.; Berzofsky, J.A.; DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **1987**, *195*, 659–685. [CrossRef]
28. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook*; Walker, J.M., Ed.; Humana Press: Totowa, NJ, USA, 2005; pp. 571–607.
29. Magnan, C.N.; Baldi, P. SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* **2014**, *30*, 2592–2597. [CrossRef] [PubMed]
30. Petersen, B.; Petersen, T.N.; Andersen, P.; Nielsen, M.; Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **2009**, *9*, 51. [CrossRef] [PubMed]
31. Magnan, C.N.; Zeller, M.C.; Kayala, M.A.; Vigil, A.; Randall, A.; Felgner, P.L.; Baldi, P. High-throughput prediction of protein antigenicity using protein microarray data. *Bioinformatics* **2010**, *26*, 2936–2943. [CrossRef] [PubMed]
32. Sweredoski, M.J.; Pierre Baldi, P. COBEpro: A novel system for predicting continuous B-cell epitopes. *Protein Eng. Des. Sel.* **2009**, *22*, 113–120. [CrossRef] [PubMed]
33. Larsen, J.E.; Lund, O.; Morten Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2006**, *2*, 2. [CrossRef] [PubMed]

34. Singh, H.; Ansari, H.R.; Raghava, P.S.G. Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS ONE* **2013**, *8*, e62216. [CrossRef] [PubMed]
35. Magnan, C.N.; Randall, A.; Baldi, P. SOLpro: Accurate sequence-based prediction of protein solubility. *Bioinformatics* **2009**, *25*, 2200–2207. [CrossRef] [PubMed]
36. Chang, C.C.; Li, C.; Webb, G.I.; Tey, B.; Song, J.; Ramanan, R.N. Periscope: quantitative prediction of soluble protein expression in the periplasm of *Escherichia coli*. *Sci. Rep.* **2016**, *6*, 21844. [CrossRef] [PubMed]
37. Kelley, L.A.; Mezulis, S.; Yates, C.M.; Wass, M.N.; Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **2015**, *10*, 845–858. [CrossRef] [PubMed]
38. Wass, M.N.; Kelley, L.A.; Sternberg, M.J. 3DLigandSite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* **2010**, *38*, W469–W473. [CrossRef] [PubMed]
39. Wass, M.N.; Sternberg, M.J. Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins* **2009**, *77* (Suppl. S9), 147–151. [CrossRef] [PubMed]
40. Veljkovic, N.; Glisic, S.; Prljic, J.; Perovic, V.; Botta, M.; Veljkovic, V. Discovery of new therapeutic targets by the informational spectrum method. *Curr. Protein Pept. Sci.* **2008**, *9*, 493–506. [CrossRef] [PubMed]
41. Tintori, C.; Manetti, F.; Veljkovic, N.; Perovic, V.; Vercammen, J.; Hayes, S.; Massa, S.; Witvrouw, M.; Debyser, Z.; Veljkovic, V.; et al. Novel virtual screening protocol based on the combined use of molecular modeling and electron-ion interaction potential techniques to design HIV-1 integrase inhibitors. *J. Chem. Inf. Model.* **2007**, *47*, 1536–1544. [CrossRef] [PubMed]
42. Cosic, I. *The Resonant Recognition Model of Macromolecular Bioactivity: Theory and Applications*; Birkhäuser: Basel, Switzerland, 1997; pp. 1–87.
43. Veljkovic, V.; Veljkovic, N.; Esté, J.A.; Hüther, A.; Dietrich, U. Application of the EIIP/ISM bioinformatics concept in development of new drugs. *Curr. Med. Chem.* **2007**, *14*, 441–453. [CrossRef] [PubMed]
44. Štambuk, N.; Manojlović, Z.; Turčić, P.; Martinić, R.; Konjevoda, P.; Weitner, T.; Wardega, P.; Gabričević, M. A simple three-step method for design and affinity testing of new antisense peptides: An example of Erythropoietin. *Int. J. Mol. Sci.* **2014**, *15*, 9209–9223. [CrossRef] [PubMed]
45. Eisenberg, D.; Weiss, R.M.; Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 140–144. [CrossRef] [PubMed]
46. Cornette, J.L.; Margalit, H.; Berzofsky, J.A.; DeLisi, C. Periodic variation in side-chain polarities of T-cell antigenic peptides correlates with their structure and activity. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 8368–8372. [CrossRef] [PubMed]
47. Quinn, T.P.; Tweedy, N.B.; Williams, R.W.; Richardson, J.S.; Richardson, D.C. Betadoublet: De novo design, synthesis, and characterization of a 3-sandwich protein. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 8747–8751. [CrossRef] [PubMed]
48. Schneider, J.P.; Pochan, D.J.; Ozbas, B.; Rajagopal, K.; Pakstis, L.; Kretsinger, J. Responsive hydrogels from the intramolecular folding and self-assembly of a designed peptide. *J. Am. Chem. Soc.* **2002**, *124*, 15030–15037. [CrossRef] [PubMed]
49. Griffioen, A.W.; van der Schaft, D.V.J.; Barendsz-Janson, A.F.; Cox, A.; Boudier, H.A.J.S.; Hillen, H.F.P.; Mayo, K.H. Anginex, a designed peptide that inhibits angiogenesis biochem. *Biochem. J.* **2001**, *354*, 233–242. [CrossRef] [PubMed]
50. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [CrossRef]
51. Brown, F.; Doughan, G.; Hoey, E.M.; Martin, S.J.; Rima, B.K.; Trudgett, A. *Vaccine Design*; Wiley: Chichester, UK, 1993; pp. 33–44.
52. Cheng, J.; Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* **2006**, *22*, 1456–1463. [CrossRef] [PubMed]
53. Ratanji, K.D.; Derrick, J.P.; Rebecca, J.; Dearman, R.J.; Kimber, I. Immunogenicity of therapeutic proteins: Influence of aggregation. *J. Immunotoxicol.* **2014**, *11*, 99–109. [CrossRef] [PubMed]
54. Patel, S.C.; Bradley, L.H.; Jinadasa, S.P.; Hecht, M.H. Cofactor binding and enzymatic activity in an unevolved superfamily of de novo designed 4-helix bundle proteins. *Protein Sci.* **2009**, *18*, 1388–1400. [CrossRef] [PubMed]
55. Patel, S.C.; Hecht, M.H. Directed evolution of the peroxidase activity of a de novo-designed protein. *Protein Eng. Des. Sel.* **2012**, *25*, 445–452. [CrossRef] [PubMed]

56. Shi, Y.; Shi, Y. Metabolic enzymes and coenzymes in transcription—A direct link between metabolism and transcription? *Trends Genet.* **2004**, *20*, 445–452. [CrossRef] [PubMed]
57. Naqvi, A.A.T.; Hassan, I. Design, principles, network architecture and their analysis strategies as applied to biological systems. In *Systems Biology Application in Synthetic Biology*; Singh, S., Ed.; Springer: New Delhi, India, 2016; pp. 21–31.
58. Perovic, V.R.; Muller, C.P.; Niman, H.L.; Veljkovic, N.; Dietrich, U.; Tosic, D.D.; Glisic, S.; Veljkovic, V. Novel phylogenetic algorithm to monitor human tropism in Egyptian H5N1-HPAIV reveals evolution toward efficient human-to-human transmission. *PLoS ONE* **2013**, *8*, e61572. [CrossRef] [PubMed]
59. Dereeper, A.; Guignon, V.; Blanc, G.; Audic, S.; Buffet, S.; Chevenet, F.; Dufayard, J.-F.; Guindon, S.; Lefort, V.; Lescot, M.; et al. Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **2008**, *36*, W465–W469. [CrossRef] [PubMed]
60. Dereeper, A.; Audic, S.; Claverie, J.-M.; Blanc, G. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evolut. Biol.* **2010**, *10*, 8. [CrossRef] [PubMed]
61. Cicchillo, R.M.; Baker, M.A.; Schnitzer, E.J.; Newman, E.B.; Krebs, C.; Booker, S.J. *Escherichia coli* L-serine deaminase requires a [4Fe-4S] cluster in catalysis. *J. Biol. Chem.* **2004**, *279*, 32418–32425. [CrossRef] [PubMed]
62. Digianantonio, K.M.; Korolev, M.; Hecht, M.H. A non-natural protein rescues cells deleted for a key enzyme in central metabolism. *ACS Synth. Biol.* **2017**. [CrossRef] [PubMed]
63. Young, D. *Computational Drug Design*, 1st ed.; Wiley: Hoboken, NJ, USA, 2009; pp. 1–5.
64. Higgs, P.G.; Attwood, T.K. *Bioinformatics and Molecular Evolution*, 1st ed.; Blackwell: Malden, MA, USA, 2005; pp. 1–10.
65. Petoukhov, S.V. The system-resonance approach in modeling genetic structures. *Biosystems* **2016**, *139*, 1–11. [CrossRef] [PubMed]
66. Jiménez-Montano, M.A.; Coronel-Brizio, H.F.; Hernández-Montoya, A.R.; Ramos-Fernández, A. Codon information value and codon transition-probability distributions in short-term evolution. *Physica A* **2016**, *454*, 117–128. [CrossRef]
67. StatSoft, Inc. STATISTICA (Data Analysis Software System), Version 7. 2014. Available online: http://www.statsoft.com (accessed on 31 July 2016).
68. Hammer, Ø.; Harper, D.A.T.; Ryan, P.D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontol. Electron.* **2001**, *4*, 9.
69. Hocke, K.; Kämpfer, N. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. *Atmos. Chem. Phys.* **2009**, *9*, 4197–4206. [CrossRef]
70. Zhao, W.; Agyepong, K.; Serpedin, E.; Dougherty, E.R. Detecting periodic genes from irregularly sampled gene expressions: A comparison study. *EURASIP J. Bioinform. Syst. Biol.* **2008**, 769293. [CrossRef] [PubMed]
71. Okonechnikov, K.; Golosova, O.; Fursov, M.; UGENE Team. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **2012**, *28*, 1166–1167. [CrossRef] [PubMed]
72. Štambuk, N.; Konjevoda, P.; Manojlović, Z.; Štambuk, A.; Turčić, P.; Gotovac, N. Synthetic proteins designed using ternary coding patterns: From nucleotide information to protein structure, function and music. *Symmetry Cult. Sci.* **2016**, *27*, 163–171.