Analysis of Coiled Coil Oligomerization - A Multidisciplinary Approach

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Dipl. Biol. CARSTEN C. MAHRENHOLZ, MBA

born in Wiesbaden

Berlin, September 2010

Research for this thesis was conducted from 2007 to 2010,
supervised by Dr. Rudolf Volkmer
at the *Molecular Libraries Group, Charité Berlin.*

1st Reviewer: Prof. Dr. Hans-Dieter Volk, *Institute of Medical Immunology, Charité Berlin*

2nd Reviewer: Prof. Dr. Beate Koksch, *Institute for Chemistry and Biochemistry, Freie Universität Berlin*

date of defence December the 14th, 2010

This work was generously funded by the *Manchot Foundation (Henkel KGaA)* and supported by scholar- and fellowships from the *Federation of the Societies of Biochemistry and Molecular Biology*, the *Charité Medical School*, and *GlaxoSmithKline*.

This work would not have been possible without the kind support of: Rudolf Volkmer, Christiane Landgraf, Ines Kretzschmar, Victor Tapia, and all group members of the *Molecular Libraries, Charité, Berlin*; Ulrich Bodenhofer, Sepp Hochreiter, *Johannes Kepler University, Linz.*; Beate Koksch, *Freie Universität, Berlin;* Sandro Keller, *Leibnitz Institute for Molecular Pharmacology, Berlin*; Michel Steinmetz, *Paul Scherrer Institute, Villigen*; Holger Strauss, *Novo Nordisk, Kopenhagen* and Nikolaus Ernsting, *Humboldt University, Berlin*. Victor Greiff, Johannes Eckstein, Michal Or'Guil, Prisca Boisguérin, and Nicole Wittenbring contributed valuable comments on the manuscript.

My special thanks go to Ingrid Abfalter for moral support and many fruitful and stimulating discussions that influenced this work.

Abstract in english

Understanding the relationship between protein sequence and structure is one of the great challenges in biology. Since in the case of the ubiquitous coiled coil motif, structure and occurrence have been described in extensive detail, it might stand to reason that we have a clearly drawn picture of coiled coils. However, the rules for oligomeric formation, and thus the key to biological function, are poorly understood.

This work investigates the oligomerization of coiled coils by means of a multidisciplinary approach that combines biochemistry, biophysics, and bioinformatics to shed new light on the formation of two- and three-stranded coiled coils:

Based on comprehensive peptide libraries of GCN4 and other coiled coil mutants, the influence of amino acid substitutions on their association is examined. Furthermore, this work uses a machine learning approach to tackle coiled coil oligomerization and identify its underlying rules in the form of weighted amino acid patterns. These rules form the basis of the highly reliable classification tool PrOCoil, which also visualizes the contribution of each individual amino acid to the overall oligomeric tendency of a given coiled coil sequence.

Thus, for the first time, a complete network of sequence parameters that influence oligomerization is established, and the underlying rules of coiled-coil formation are presented.

This work is rounded off by a methodical contribution. In order for a method to provide a basis for drawing sound conclusions, it must be reviewed carefully. In the case of peptide libraries, little is known about the cross-reactivity between peptides and detection agents. A systematic review and appraisal of the potential of three common read-out systems – 5(6)-TAMRA, FITC, and biotin/streptavidin-POD – to cross-react with individual amino acids in a peptide sequence is therefore presented.

Abstract in german

Das Verständnis der Beziehung zwischen Sequenz und Struktur von Proteinen ist eine der großen Herausforderungen der heutigen Biologie. Im Falle des weit verbreiteten Coiled-Coil-Motivs sind speziell Struktur und Vorkommen detailliert beschrieben. Es ist also naheliegend, von einer vollständig aufgeklärten Struktur auszugehen. Um so erstaunlicher ist aber, dass die Coiled-Coil-Oligomerisierung – zentrales Kriterium für die biologische Funktion dieser Proteine – nahezu unverstanden ist.

In dieser Arbeit wird das Phänomen der Coiled-Coil-Oligomerisierung anhand eines multidisziplinären Ansatzes untersucht. Erst die Kombination aus Biochemie, Biophysik und Bioinformatik erlaubt es, die Formation von zwei- und dreisträngigen Coiled-Coils zu erklären:

Zu diesem Zweck wird auf Basis von umfangreichen Peptidbibliotheken von GCN4 und anderen Coiled-Coil-Mutanten der Einfluss von Aminosäure-Substitutionen auf das Assoziationsverhalten untersucht. Weiterhin beschäftigt sich die vorliegende Arbeit mit der Untersuchung des Oligomerisierungsverhaltens von Coiled-Coils. Basierend auf einer neuen Theorie und unter Zuhilfenahme von Support Vector Maschinen werden die der Oligomerisierung zugrundeliegenden Regeln präsentiert. Diese Regeln, in Form von gewichteten Beziehungen zwischen Aminosäuren, bilden die Grundlage eines neuartigen Klassifikations-Tools. "PrOCoil" ist in der Lage, die Stöchiometrie von Coiled-Coils mit außergewöhnlicher Genauigkeit vorherzusagen und den Beitrag einzelner Aminosäuren dazu zu visualisieren. In Form eines Netzwerks von Sequenzparametern wird hier erstmalig ein Modell eingeführt, das in der Lage ist, die Coiled-Coil Oligomerisierung zu erklären.

Aus methodischer Sicht feit die Anwendung einer Standard-Methode nicht vor kritischer Reflexion. Unabdingbar für eine zuverlässige Interpretation von Peptidbibliotheken ist das Wissen um potenzielle Kreuzreaktivität von membrangebundenen Peptiden mit den Nachweisreagenzien des Analyten. Daher beinhaltet diese Arbeit als dritten Focus eine Begutachtung und Bewertung von drei in diesem Zusammenhang häufig genutzten Nachweissystemen. 5(6)-TAMRA, FITC und Biotin/Streptavidin-POD werden auf ihre Kreuzreaktivität mit einzelnen Aminosäuren in Peptidsequenzen hin untersucht.

Contents

Abbreviations

BLAST	basic local alignment search tool
BLOSUM	blocks substitution matrix
CCD	camera charging device
ELISA	enzyme-linked immunosorbent assay
FDR	false discovery rate
h	hydrophobic
HIV	human immunodeficiency virus
HPLC	high-performance liquid chromatography
LIBSVM	library for support vector machines
MALDI-TOF	matrix-assisted laser desorption ionization - time of flight
MW	molecular weight
PDB	protein data bank
POD	peroxidase
PSV	partial specific volume
PSVM	potential support vector machine
SI	signal intensity
SPR	surface plasmon resonance
SVM	support vector machine
TBS	tris-buffered saline
wt	wildtype

Chemical abbreviations

Boc	*tert*-butyloxycarbonyl
DIC	diisopropylcarbodiimide
DMF	dimethylformamide
DMSO	dimethylsulfoxide
FITC	fluorescein-substituted thiourea derivative
Fmoc	(9-fluorenyl)methoxycarbony
HOBt	1-hydroxybenzotriazole
NMP	1-methyl-2-pyrrolidone
OPfp	pentafluorophenyl ester
OtBu	*tert*-butyl ester
Pbf	2, 2, 4, 6, 7-pentamethyldihydrobenzofuran-5-sulfonyl
PyBOP	benzotriazolyloxy-tris(pyrrolidino)phosphonium hexafluorophosphate
TAMRA	5-(and 6)-carboxytetramethylrhodamine
TBTU	O-(1*H*-benzotriazol-1-yl)-N,N,N',N'-tetramethyluroniumtetrafluoroborate
tBu	*tert*-butyl
TFA	trifluoro-acetic acid
Trt	trityl

Amino acids

Ala	A	alanine	Leu	L	leucine
Arg	R	arginine	Lys	K	lysine
Asn	N	asparagine	Met	M	methionine
Asp	D	aspartic acid	Phe	F	phenylalanine
Cys	C	cysteine	Pro	P	proline
Gln	Q	glutamine	Ser	S	serine
Glu	E	glutamic acid	Thr	T	threonine
Gly	G	glycine	Trp	W	tryptophan
His	H	histidine	Tyr	Y	tyrosine
Ile	I	isoleucine	Val	V	valine

Introduction

Coiled coils | 1.1

In 1952, L. Pauling [Pauling *et al.*, 1953] and F. H. C. Crick [Crick, 1952] first described the structure of the α-helical coiled coil. Since then it has become a prototypical textbook example of a structural motif, being commonly described as consisting of between two and seven α-helices. Almost 6% of the proteins in the Protein Data Bank (PDB) [Bernstein *et al.*, 1977] contain coiled coil regions [Hadley *et al.*, 2008], of which more than 90% show dimeric or trimeric interactions. Due to their ability to oligomerize, coiled coils perform, either on their own or as part of larger protein complexes, a variety of important cellular functions [Burkhard *et al.*, 2001]. Their ubiquity and the stable interactions of their helices make coiled coils ideal building blocks for designing novel proteins. Furthermore, coiled coil interactions have recently attracted attention as promising drug targets [Strauss *et al.*, 2008]. Their use in successful inhibition of membrane fusion proteins of viruses such as HIV [Bianchi *et al.*, 2005] and avian influenza [Russell *et al.*, 2008] supports the concept of rational drug design based on coiled coil proteins [McFarlane *et al.*, 2009]. Nowadays, coiled coils are used extensively and successfully to rationally design multistranded structures for applications including basic research, biotechnology, nanotechnology, material science, and medicine. The wide range of applications and the important functions these structures play in almost all biological processes highlight the need for a detailed understanding of the factors that control coiled-coil folding and oligomerization.

Today, a plethora of information about coiled coils is available, including their prevalence, sequence characteristics, and structures. As illustrated in **Figure 1**, they have in common a periodically recurrent sequence called a heptad repeat of the form $(abcdefg)_n$. Usually, the positions *a* and *d* in these repeats are occupied by hydrophobic amino acids located at the hydrophobic core crucial for tertiary structure, while positions *e* and *g* typically are charged residues [O'Shea *et al.*, 1993].

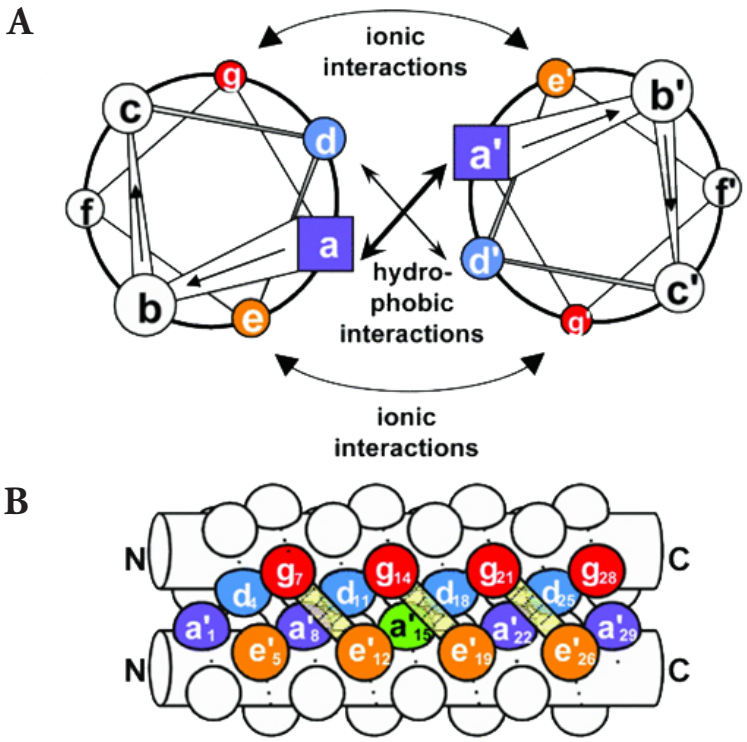


Figure 1 | Schematic representation and structure of the parallel dimeric coiled-coil motif. (A) Helical wheel diagram looking down the helical axis from the N- to the C-terminus. Heptad positions are labeled *a*–*g* and *a'*–*g'* respectively. Positions *a*, *d*, *e*, and *g* are color-coded. **(B)** In the side view, the helical backbones are represented by cylinders, the side chains by knobs, and the path of the polypeptide chain is indicated by a line wrapped around the cylinders. For simplicity, the supercoiling of the helices is not shown. While residues at positions *a* (purple) and *d* (blue) make up the hydrophobic interface, residues at positions *e* (orange) and *g* (red) pack against the hydrophobic core. They can participate in interhelical electrostatic interactions between residue i (*g* position) of one helix and residue i'+5 of the other helix (*e'* position in the next heptad), as indicated by the hatched bars. Also indicated is the core *a* position (green), which is often occupied by polar residues mediating strand-pairing specificity. (Figure adapted from Mason *et al.* [2008])

These obvious regularities are used to predict coiled coil segments in amino acid sequences [Lupas *et al.*, 1991; Delorenzi *et al.*, 2002; McDonnell *et al.*, 2006]. Hence, one might expect our understanding of coiled coils to be complete. Most remarkably, however, the hidden and more complex rules for oligomeric formation, and thus the key to biological function, are poorly understood. A first but crude indicator of whether the oligomeric state of a coiled coil is dimeric (**Figure 2**) or trimeric (**Figure 3**) may be its intra- and extra-cellular prevalence [Lupas *et al.*, 2005], but it clearly does not provide any information about the sequence features that govern oligomerization.

Figure 2 | Examples and helical wheel diagram of dimeric coiled coils. Displayed are both the complete proteins and details of the a-helices of **(A)** tropomyosin and **(C)** the DNA-bound c-Jun/c-Jun AP1 dimer. Tropomyosin mediates the interactions between the troponin complex and actin so as to regulate muscle contraction [Lewis *et al.*, 1980]. The c-Jun oncoprotein is a major component of the transcription factor complex AP-1, which regulates the expression of multiple genes essential for cell proliferation, differentiation, and apoptosis [Hartl *et al.*, 2003]. **(B)** Helical wheel diagram of a dimer looking down the helical axis from the N- to the C-terminus. Heptad positions are labeled from a to g. Arrows represent the interhelical electrostatic interactions.

Despite extensive experimental and computational efforts such as mutation analysis, NMR, X-ray crystallography, and statistics [Lupas *et al.*, 1991; Portwich *et al.*, 2007; Gingras *et al.*, 2008; Sheriff *et al.*, 1994; Hu *et al.*, 1993], our knowledge of which oligomer a specific coiled coil forms has, until now, been limited to describing the phenomenon on the basis of a small number of protein samples.

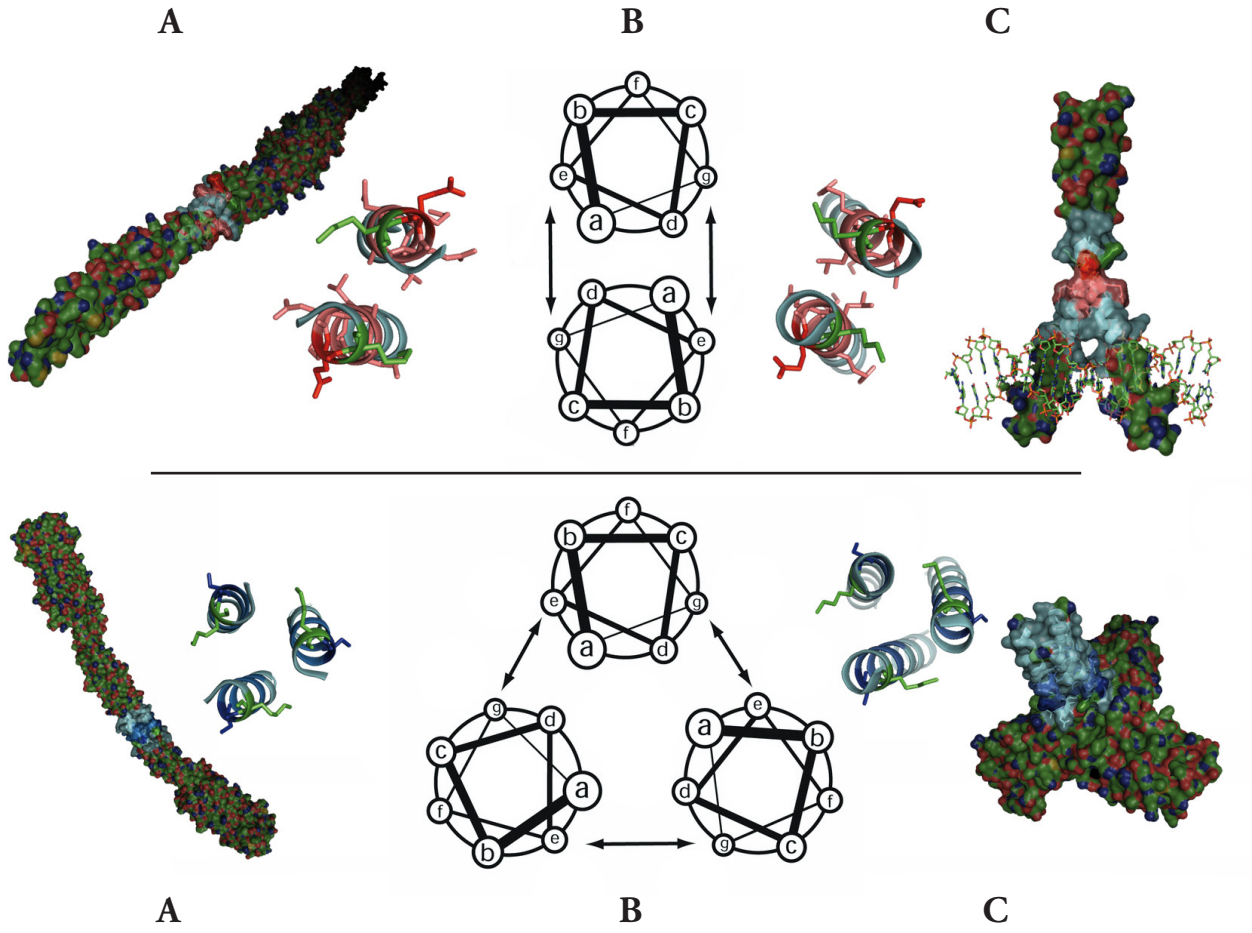


Figure 3 | Examples and helical wheel diagram of trimeric coiled coils. Displayed are both the complete proteins and details of the α-helices of **(A)** the surface transmembrane glycoprotein (GP2) of the ebola virus and **(C)** mannose-binding lectin (MBL). GP2 is responsible for binding to target cells and subsequent fusion of the viral and host-cell membranes [Malashkevich *et al.*, 1999]. MBL is a calcium-dependent serum protein that plays a role in the innate immune response by binding to carbohydrates on the surface of a wide range of pathogens [Turner, 1998]. All figures were created using PyMOL (http://pymol.sourceforge.net/) **(B)** Helical wheel diagram of a trimer looking down the helical axis from the N- to the C-terminus. Heptad positions are labeled from a to g. Arrows represent the interhelical electrostatic interactions.

Now, as the amount of available post-genomic sequence data is growing rapidly, the challenge is to explain coiled coil oligomerization by extracting an actual set of rules from this data.

Pharmacological potential | 1.1.2

Due to their structural simplicity, stability, and specificity, the leucine zipper and other coiled coil proteins have attracted attention in the context of pharmacological applications. As has been shown for the example of the oncosuppressor p53, the structure and function of a self-associating target protein in the cell can be disrupted by administering a chimeric coiled coil [Contegno *et al.,* 2002]. Another possible medical application of coiled coils is in biomaterials designed for use in the human body. For instance, it has been shown that they can form a reversibly contracting "smart" hydrogel that consists only of polypeptides [Petka *et al.,* 1998]. When adding coiled coils to a water-soluble polymer network, directed coiled coil formation results in contraction of the hydrogel and a volume reduction of up to 90% [Tang *et al.,* 2001; Wang *et al.,* 1999; Wang *et al.,* 2001]. Alternatively, the specificity of coiled coil domains in hetero-dimerization and -oligomerization can be used in controlled drug delivery and release systems [Moll *et al.,* 2001]. Both a therapeutic agent (e.g., a radionuclide-chelate complex [Goldenberg, 2003]) and a targeting component (e.g., an antibody) are attached to a hetero-dimerizing domain. First, the targeting component is administered to the organism to specifically recognize and mark the target site. Subsequently, the therapeutic agent is added, which dimerizes with the coiled coil domain of the targeting component and thus delivers the drug directly to its target site. Such two-step approaches can enhance the therapeutic effect of drugs and can reduce toxicity in non-affected tissue [Goodwin *et al.*, 2001; Knox *et al.,* 2000]. Hetero-dimerizing leucine zipper domains have several advantages over other established systems. For instance, the disadvantages of the biotin/streptavidin system are that streptavidin may provoke immune response and that it may be bound by endogenous biotin. It has been shown repeatedly both by means of combinatorial approaches and *in vivo* experiments that proteins coupled to coiled coil domains can hetero-dimerize [Behncken *et al.,* 2000; Ghosh *et al.,* 2000; Katz *et al.,* 1998]. Since infections and other causes of illness (e.g., of Diabetes mellitus [Hua *et al.,* 2000]) are linked to changes in α-helical coiled coil structures, there is a demand for agents that can easily detect these changes and/or provide a means of therapy.

Bioinformatics | 1.2

Comparing trimeric coiled coil sequences by hand, experimentalists have recently discovered the first complex trimerization pattern [Kammerer *et al.*, 2005]. However, experimental approaches can never be exhaustive, making bioinformatics the method of choice [Lupas, 2008] for identifying the underlying oligomerization rules embedded in the sequence data. Previous coiled coil research, including bioinformatics approaches, has been based on the notion that strategies using sequence homology and single amino acid distributions are adequate methods for explaining oligomerization.

State-of-the-art prediction | 1.2.1

The aforementioned heptad periodicity of coiled coils and the clear and simple appearance of their structures have made possible a large number of computational approaches to their analysis. These range from (i) simple sequence-based approaches, counting single and pairwise residue distributions, to (ii) approaches based on Hidden Markov Models without scanning windows, (iii) structure-based approaches, detecting knobs-into-holes packing in helical bundles, and (iv) approaches based on matrices of residue frequencies that aim to distinguish different oligomeric tendencies.

The earliest sequence-based approaches were initially used to detect periodicities in the basic heptad pattern [see e.g. McLachlan *et al.*, 1983], but subsequently also for detecting deviations from the heptad pattern itself [Hoiczyk *et al.*, 2000]. Based on residue distributions at each of the seven heptad positions of the putative coiled coil segments of myosin, tropomyosin, a-keratin, and hemagglutinin, a second widely used sequence-based approach was implemented in the form of the prediction tool Coils (www.ch.embnet.org/software/ COILS_form.html). It uses scanning-window-based residue frequencies to predict whether a sequence of unknown structure forms a coiled coil [Lupas *et al.*, 1991]. A variant of this approach, using pairwise residue correlations, was

developed and implemented by Berger and colleagues in the program PairCoil (paircoil.lcs.mit.edu/cgibin/paircoil) [Berger *et al.*, 1995]; LearnCoil, a further variant, can be trained iteratively on a set of target proteins [Berger *et al.*, 1997].

The most promising sequence-based approach, MARCOIL (www.isrec.isbsib.ch/BCF/Delorenzi/Marcoil/index.html), builds upon Hidden Markov models [Delorenzi *et al.*, 2002] and operates without a scanning window, thus removing a limitation of Coils and PairCoil.

The main structure-based program used for the analysis of coiled coils is SOCKET (www.biols.susx.ac.uk/Biochem/Woolfson/html/coiledcoils/socket/) [Walshaw *et al.*, 2001]. This tool was designed to detect knobs-into-holes packing in helical bundles and represents the most direct way of evaluating the compatibility of a structure with the standard model. The program operates by representing side-chains by their centers of mass and classifying them as knobs if they contact four or more side-chain centers within a specified distance cutoff (set to 7.0 Å by default). At the same time, the program assigns an orientation, a register, and the number of constituent helices for each detected coiled coil.

In order to discriminate between two- and three-stranded coiled coils, matrices of residue frequencies have been used with fair success. Building upon such matrices, Woolfson and Alber [1995] developed the program Scorer, and Wolf *et al.* [1997] the program MultiCoil (multicoil.lcs.mit.edu/cgibin/multicoil).

Support vector machines | 1.2.2

In recent years, SVMs have become established as a standard tool in machine learning, and their popularity for biosequence classification has increased dramatically [Schölkopf *et al.*, 2004]. SVMs provide mathematically sound classifications even if the dataset is too small to achieve significant results with probabilistic techniques [Vapnik, 1998]. In fact, SVMs are the method of choice both because they can be used to distinguish dimers from trimers and because, at the same time, they

also provide the rules (weighted patterns) on which their decisions are based and which are so valuable for protein design purposes.

In the context of classifying biological sequences as described here, SVMs require a kernel that obtains two sequences as input and supplies a scalar value as a measure for their similarity. By extensive testing in cooperation with the Institute of Bioinformatics, JKU Linz, it was verified that the currently most popular sequence kernels (spectrum [Leslie *et al.*, 2002] and mismatch kernel [Leslie *et al.*, 2003]) are not the best choice for classifying coiled coils, as they measure similarity by counting long, continuous substrings shared by both test and query sequences. Therefore, inspired by earlier approaches that make use of pairwise residue co-occurrences [Berger *et al.*, 1995; Wolf *et al.*, 1997; Fong *et al.*, 2004; McDonnell *et al.*, 2006], and motivated by the findings in this work, a new kernel – the coiled coil kernel – was developed. In contrast to the substrings used in other sequence kernels, the pairs of amino acids used by the coiled coil kernel need not be adjacent [Berger *et al.*, 1995]. Using the coiled coil kernel, a SVM generates rules by optimizing the pattern weights such that the combined rules achieve maximum discrimination between dimers and trimers. For a more detailed description of the approach, see the Methods section of this work.

SPOT-synthesis | 1.3

The growing demand for binding assays to study protein-protein interaction can be addressed by peptide array-based methods. The SPOT technique is a widespread peptide-array technology, which is able to distinguish semi-quantitatively the binding affinities of peptides to defined protein targets within one array.

Introduced in 1963 by Robert Bruce Merrifield, solid phase synthesis [Merrifield *et al.*, 1963] became an all-important technique for synthesizing peptides and small protein domains. The potential of this automated method is limited only by the fact that it is not feasible to synthesize and screen large numbers of peptides. Motivated by the great demand for rapid and effective parallel synthesis of peptides, in 1992 Ronald Frank developed and published a method for synthesizing peptides by spot-wise coupling of small amounts of activated amino acids directly on a membrane and for subsequent screening of large peptide arrays on these planar cellulose supports [Frank, 1992].

While SPOT synthesis is a positionally addressable, highly parallel, and technically simple experimental procedure, it is also a very flexible and economic technique that can be applied to a broad spectrum of biological tasks [Reinecke *et al.*, 2001; Frank, 2002]. The special properties of solid phase membrane supports – made, for instance, of cellulose or other polymers – allow screening methods such as binding assays, enzymatic assays, and cellular assays [Wenschuh *et al.*, 2000]. The planar membrane materials must be compatible with reagents, solvent systems, and reaction conditions (chemical compatibility) as well as resistant to washing and cleavage operations (mechanical stability). In most cases, planar cellulose membrane sheets are used for the synthesis and subsequent screening of peptide arrays.

The SPOT technique uses conventional Fmoc chemistry, and the general concept for SPOT synthesis is summarized as follows: (i) membrane functionalization; (ii) spacer and/or linker attachment; (iii) positionally addressed SPOT synthesis; (iv) cleavage, i.e., side chain deprotection; (v^a) solid-phase screening of membrane-bound peptides; or (v^b) cleavage from the membrane for solution-phase assays. For a schematic overview

of the coupling and deprotection cycle see **Figure 4**.

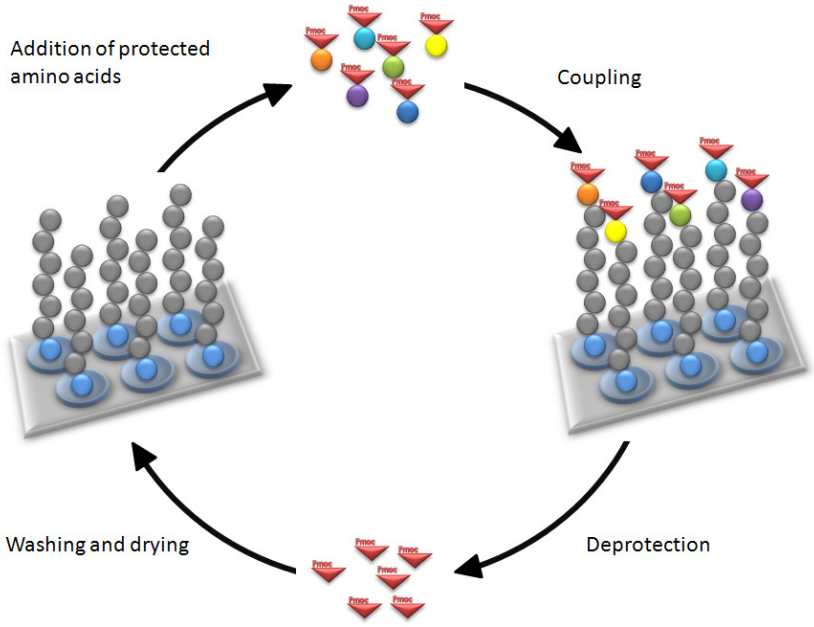


Figure 4 | Schematic illustration of the coupling cycle in SPOT synthesis. The processes of (i) addition of protected amino acids, (ii) coupling, (iii) deprotection, and (iv) washing are repeated until peptides of the desired length are bound to the membrane. For further information see the Methods section.

Measuring protein-protein interactions | 1.3.1

There are several commonly used methods for measuring protein-protein interactions and binding affinities, such as enzyme-linked immunosorbent assay (ELISA) and surface plasmon resonance (SPR). In contrast to most of these methods, protein and peptide arrays on planar surfaces [Frank, 2002; Andresen *et al.*, 2009; Tao *et al.*, 2007; Stoevesandt *et al.*, 2009] allow high-throughput measurement because they provide a higher density of probes, so a multitude of molecular interactions can be measured in parallel. Array experiments have demonstrated their value for bimolecular binding assays [Pritchard *et al.*, 2008; Li *et al.*, 2002], especially in the case of protein-protein interactions [Stoevesandt *et al.*, 2009; Zhu *et al.*, 2001; Beutling *et al.*, 2008; Volkmer *et al.*, 2009].

Synthetic peptide arrays have several advantages [Volkmer *et al.*, 2009]: (i) peptide synthesis is faster and cheaper than expression-related techniques, (ii) peptide probes are stable moieties, and (iii) peptide synthesis allows incorporation of non-gene-encoded residues. A drawback of

applying peptides instead of whole proteins as probes is information loss due to the missing structural context. This can be compensated for by adapting the task of the peptide array experiment, for example, by focusing on modular binding events or by resolving immunorecognition to the epitope level. Peptide arrays are usually prepared in a micro- or macro-array format [Reimer *et al.*, 2002]. The latter kind of array is generated according to the SPOT synthesis approach [Frank, 1992], which is accessible even for non-specialized laboratories. The SPOT technology and many of its applications have been reviewed extensively [e.g. Frank, 2002, Beutling *et al.*, 2008; Volkmer, 2009]. In principle, signal intensities – the output of this technique – can be used to roughly distinguish between different affinities [Weiser *et al.*, 2005]. The most important application of the SPOT technique, however, is to differentiate qualitatively between binding affinities of peptides to defined protein targets within one array, using fluorescent or chemiluminescent read-out systems.

Interference of screening systems | 1.3.2

The quality of an assay system used for probing peptide arrays depends on the well balanced combination of screening and read-out methods. The former address the steady state of analyte capture, while the latter provide the means to detect captured analyte. Usually, both screening and read-out are carried out directly on the peptide array and are often performed as separate procedures. The visualization of peptides binding the interaction partner is done in an additional step, in which the probed peptide array is subsequently immersed in a solution containing a label conjugate with high binding affinity to the analyte. Besides antibody-based immunoblotting techniques [Wilson *et al.*, 1978; Harlow *et al.*, 1988], the biotin/streptavidin-POD system has recently been reported as a convenient combination of a non-interfering screening strategy (biotin-conjugated analytes) with a specific affinity-based read-out strategy (streptavidin-conjugated reporter) for peptide arrays [Winkler *et al.*, 2008; Dürauer *et al.*, 2006]. More advantageously though, screening and read-out can be achieved simultaneously by

direct labeling of the analyte with a detectable moiety, for instance, with fluorescent dyes. These dyes can be incorporated synthetically [Toepert *et al.*, 2003; Portwich *et al.*, 2007] or via methods used in activity-based protein profiling [Uttamchandani *et al.*, 2008]. In all cases, however, false positive results can occur when challenging a peptide array with analyte or detecting captured analyte with label conjugates. This is due to the diversity of mechanisms by which peptides may interact directly with any of the detection agents. Control incubations, using only the detection agents, for the read-out procedure are obligatory and standard in good laboratory practice.

Milestones and Objectives

The aim of this work is to investigate which specific properties of a coiled coil domain influence its oligomerization by means of a multi-method approach using (a) biochemical, (b) biophysical, and (c) bioinformatics methods.

Using comprehensive peptide libraries of coiled coil mutants, the influence of amino acid substitutions on association are tested, and the association is examined further by biophysical methods. The feasibility of the biochemical methods to (i_a) distinguish coiled coils from non-coiled coils and to (ii_a) investigate the oligomeric state is evaluated. To be able to draw sound conclusions, the biochemical core method, the SPOT synthesis, must be reviewed carefully: Despite being a general problem, little is known about the cross-reactivity of peptides with the detection agents, which leads to false positive results. To this end ($iii_{a,b}$) three common agents 5-(and 6)-carboxytetramethylrhodamine (TAMRA), fluoresceinisothiocyanate in the form of the peptide-bound fluorescein-substituted thiourea derivative (FITC), biotin and streptavidin-POD are tested for cross-reaction with individual amino acids in a peptide sequence.

Bioinformatics and statistics are used to expand the knowledge of coiled coil oligomerization beyond experimental data.

Three postulates form the starting point of this thesis that aims to shed new light on the theory of coiled coil oligomerization:

I. Analysis of simple amino acid distributions is insufficient to distinguish oligomers.

II. Relationships between amino acids at different positions are important.

III. Taken together, the relationships form a network that determines structure.

(iv_c) These postulates are examined using machine learning methods and other statistical methods, which – if successful – lead to a revised theory of coiled coil oligomerization.

Based on all known dimeric and trimeric coiled coils in the Protein Data Bank (PDB), pairwise feature extraction is used to ($v_{a,c}$) identify characteristic amino acid patterns that determine the rules for oligomerization. These rules are used as a basis for a classification tool that makes it possible to (vi_c) visualize the contribution of each individual amino acid to the overall oligomeric tendency of a given coiled coil sequence. These results are ($vii_{b,c}$) verified both computationally and experimentally, and used to ($viii_{a,c}$) explain the hitherto puzzling behavior of the yeast transcriptional activator GCN4, which can, as a result of minimal mutations in its amino acid sequence, switch from forming a dimer to forming a trimer [Portwich *et al.*, 2007].

Thus, this work tackles three **central tasks**:

- Analysis of three common detection agents (TAMRA, FITC and biotin/streptavidin-POD) to reveal their ability to cross-react with cellulose-bound peptides.

- Analysis of coiled coils using peptide libraries.

- Development and testing of a new sequence-based theory to explain coiled coil oligomerization.

Methods

Biochemical methods | 3.1

All reagents and solvents were ordered from Aldrich (Steinheim, Germany), Fluka (Buchs, Suisse), Merck (Darmstadt, Germany) or Sigma-Aldrich (Seelze, Germany). Other sources and companies are mentioned in the text.
The water used was desalinated and demineralized using a Simplicity 185 unit (Millipore, Billerica, USA).

SPOT-synthesis | 3.1.1

Cellulose-bound peptide arrays were prepared according to standard SPOT synthesis protocols using a SPOT synthesizer (Intavis, Köln, Germany) as described in detail in [Wenschuh *et al.*, 2000]. The peptides were synthesized on amino-functionalized cellulose membranes (Whatman, Maidstone, Great Britain) of the ester type prepared by modifying cellulose paper with Fmoc-β-alanine as the first spacer residue. In the second coupling step, the anchor position Fmoc-β-alanine-OPfp in dimethylsulfoxide (DMSO) was used. Residual amino functions between the spots were capped by acetylation. The Fmoc group

was cleaved using 20% piperidine in dimethylformamide (DMF). The cellulose-bound peptide arrays were assembled on these membranes by using 0.3 M solutions of Fmoc-amino acid-OPfp in 1-Methyl-2-pyrrolidone (NMP). Side-chain protection of the Fmoc-amino acids used was as follows: Glu, Asp (OtBu); Ser, Thr, Tyr (tBu); His, Lys, Trp (Boc); Asn, Gln, Cys (Trt); Arg (Pbf). After the last coupling step, the acid-labile protection groups of the amino acid side chains were cleaved using 90% trifluoro-acetic acid (TFA) for 30 min and 60% TFA for 3 h. Peptides were cleaved from the membrane using the standard protocol as described in [Wenschuh *et al.*, 2000] and dissolved in water (using 10% acetonitrile to increase solubility if necessary).

Analysis and purification | 3.1.2

HPLC analysis (Waters, Milford, USA) was conducted using a linear solvent gradient (A: 0.05% TFA in water; B: 0.05% TFA in acetonitrile; gradient: 5–60% B over 30 min; UV detector at 214 nm; RP-18 column).

α-cyanocinnamic acid was used as a matrix for MALDI-TOF (Applied Biosystems, Forster City, USA) MS analysis.

Peptide synthesis on resin | 3.1.3

Soluble peptides were synthesized (50 µmol scale) as amides on a multiple synthesizer according to the standard Fmoc machine protocol using TentaGel S RAM resin (Rapp Polymere, Tübingen, Germany) and PyBOP activation. Each peptide was modified N-terminally with 5-(and 6)-carboxytetramethylrhodamine (using TBTU activation), fluoresceinisothiocyanate (using HOBt, DIC activation), or biotin (using PyBOP activation). All peptides were analyzed by reversed phase HPLC and MALDI-TOF. HPLC purification and analysis were conducted as described above.

Binding studies on cellulose membranes | 3.1.4

All incubation and washing steps were carried out under gentle shaking and at room temperature. After washing the membrane with ethanol once for 10 min and three times for 10 min with Tris-buffered saline (TBS: 50 mM Tris-(hydroxymethyl)-aminomethane, 137 mM NaCl, 2.7 mM KCl, adjusted to pH 8 with HCl/0.05%), the membrane-bound peptide arrays were blocked (3 h) with blocking buffer (casein-based blocking buffer concentrate (Sigma-Genosys, Cambridge, UK), 1:10 in TBS containing 5% (w/v) sucrose), and then washed with TBS (1x10 min). Subsequently, the peptide arrays were incubated with the labeled analytes (c = 10 µM) for 10 min in TBS blocking buffer. After washing for 120 min with TBS, analysis and quantification of peptide-bound dyes/biotin/streptavidin-POD were carried out using a Lumi-Imager. For biotin/streptavidin-POD, a chemiluminescent substrate was added beforehand. For densiometric analysis, the membranes were scanned and read out directly by GeneSpotter (Microdiscovery, Berlin, Germany).

Measurement of spot signal intensities | 3.1.5

For each detection system, binding events were recorded by a cooled CCD-camera (TAMRA-fluorescence, FITC-fluorescence, and SA-linked chemiluminescence) using a Lumi-Imager (Roche, Indianapolis, USA). Additionally, TAMRA staining was also recorded by scanning in the visible light range using a HP Scanjet G3010 (Hewlett-Packard, Böblingen, Germany), resulting in a digital image file (referred to as densiometric analysis). The signal intensity (SI) of each spot was calculated by defining a spot radius that can be optimally applied to all spots in the image and taking the median value of the pixel intensity. The background signal was determined with a safety margin to each spot's circular region, and then the global background mean was subtracted from each individual spot signal. This parameter is referred to as SI. Grid-layer and SI were calculated using dedicated image analysis software: GeneSpotter has a fully automatic grid-finding routine resulting

in reproducible signal intensities. The median value of the intraspot distribution was sufficient to avoid saturation. Results are shown as the interspot global background-corrected mean value over three replica spots for each sequence. TAMRA was measured at 645 nm, FITC at 520 nm, and streptavidin-POD via chemiluminescence. The aforementioned wavelength was chosen to detect TAMRA at lower background noise.

Biophysical methods | 3.2

Fluorescence spectroscopy | 3.2.1

Spot autofluorescence was monitored by placing freshly prepared membrane probes with peptides bound (sample) and a functionalized membrane without peptides (control) in a SPEX Flourolog 212 fluorometer (SPEX Industries, Edison, NJ) thermostated at 25°C. Excitation spectra (λ_x) were measured with the excitation wavelength varying from 260 nm to 500 nm and the emission wavelength (λ_m) set to 520 nm. Intensity of fluorescence was expressed as counts/sec, and the integration time (S/R) was 1 sec.

Circular-dichroism spectroscopy | 3.2.2

Peptides were measured in TBS (154 mM NaF, 10 mM tris(hydroxymethyl)aminomethane, pH 8.0) at room temperature. Spectra were recorded on a J-720 spectropolarimeter (Jasco, Tokyo, Japan) at a total peptide concentration of 75 µM, and corrected by subtracting the buffer baseline.

Analytical ultracentrifugation | 3.2.3

Sedimentation equilibrium experiments were conducted in a Xl-I analytical ultracentrifuge (BeckmanCoulter, Brea, USA) at 20 °C using the interference optics of the instrument. Peptide solutions were brought to dialysis equilibrium with 50 mM Tris buffer, pH 8.0, containing 137 mM NaCl and 2.7 mM KCl with PD10 columns (Amersham Bioscience, Freiburg, Germany). Peptide stock solutions had a concentration of 3 mg mL^{-1} of total peptide. 250 µL of three

different solutions (1:1, 1:2 and 1:10) were loaded into artificial-boundary centerpieces and spun at 50 or 60 krpm, depending on the expected molecular weight. Attainment of apparent sedimentation and chemical equilibrium was ascertained by comparing consecutive scans using MATCH (available from ftp://rasmb.bbri.org). Blank-corrected scans were first analyzed by calculating point-average molecular weights from the slope of a plot of the natural logarithm of the concentration versus the squared distance from the rotor axis. Visual inspection of the overlay of local molecular weights versus concentration for the different loading concentrations was used to check for reversibility and suggested the existence of limiting species at the lower and upper ends of the concentration scale. Where appropriate, this information was used to select distinct models to describe the data. These models were then fitted directly to the data using NONLIN [Johnson *et al.*, 1981]. A model was judged to be an adequate description of the data if the residuals were random by visual inspection and if no other model explained the data significantly better as judged by the variances of the fits [Otte *et al.*, 2003].

The density of the buffer used and the partial specific volume (PSV) of the different peptides were measured using a DMA 5000 densitometer (Anton Paar, Graz, Austria) as described elsewhere [Kratky *et al.*, 1993]. For mixtures of two peptides, the weighted average of the individual values of the PSV was used to calculate point-average molecular weights. A conversion factor of 3.29 fringes mg^{-1} mL^{-1} was used to convert values from fringe units to molar quantities.

Bioinformatics and statistics | 3.3

Extracting dimers and trimers from the PDB | 3.3.1

In order to create a dataset of dimeric and trimeric coiled coil sequences, the program SOCKET [Walshaw *et al.*, 2001] with a packing cutoff of 7.0 Å was used to scan the PDB for knobs-into-holes packing between helices. The output was first parsed for dimeric and trimeric sequences and then divided into parallel and anti-parallel samples. The dataset of parallel dimeric and trimeric coiled coils was refined by removing identical (sub-)sequences, as they contribute no additional sequence information. Thus, a database of 385 dimeric and 92 trimeric coiled coil sequences with heptad registers assigned by SOCKET was created (for an overview see **Figure 5** and **Table A2**).

Augmenting the database with BLAST | 3.3.2

The dataset was augmented with coiled coil sequences that were not yet structurally resolved and thus not listed in the PDB. To this end, the complete amino acid chains containing coiled coil segments were retrieved from the PDB entries and the areas SOCKET had identified as coiled coils were masked. The coiled coil masking facility of BLAST was not employed, since it is based on the prediction tool Coils. Using verified 3D data to identify coiled coil segments is the more reliable choice. Those chains that provided at least 40 unmasked amino acids were then used as inputs to BLAST [Altschul *et al.*, 1990] searches in the non-redundant (NR) database. Subsequently, a strict selection process was employed to ensure that only reliable sequences were chosen from the hits and added to the dataset: First, BLAST output sequences that were less than 85% identical to the unmasked regions of the query sequences were removed. Then the remaining sequences were used as input for the program MARCOIL [Delorenzi *et al.*, 2002] to confirm that they contain coiled coil segments and to assign their heptad registers. In the

final filtering step, only sequences that reached or exceeded a coiled coil probability of 85% according to MARCOIL were selected and included in the dataset. As depicted in **Figure 5,** this resulted in a combined PDB and approved BLAST pool of 2043 dimers and 791 trimers. The reasoning behind this particular masking procedure was that using the coiled coil region itself for a BLAST search would result in cross-hits between dimers and trimers, since they can have highly similar amino acid sequences. Searching with the chain surrounding the coiled coil, on the other hand, is very likely to provide proteins of similar structure and thus also of identical oligomerization. To check whether masking helps to avoid cross-hits, a database of the full (unmasked) chains containing the dimeric or trimeric segments collected from the PDB was created, against which a BLAST search with the masked chains was run. After removing chains that contained both dimeric and trimeric regions, no cross-hits between the set of dimers and the set of trimers were observed, which proves this method for avoiding cross-hits successful.

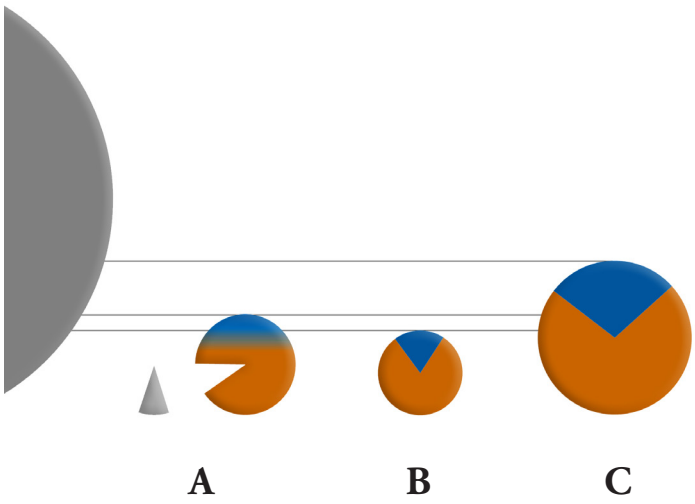


Figure 5 | Schematic overview of coiled coil datasets. Almost 6% of all proteins in the PDB (indicated on the left, grey) contain **(A)** coiled coil regions, of which more than 90% show dimeric (orange) or trimeric (blue) interaction. **(B)** Dataset of 385 dimeric and 92 trimeric structurally resolved coiled coil sequences (see Table A1)that was augmented, resulting in **(C)** a combined PDB and approved BLAST dataset of 2043 dimeric and 791 trimeric sequences. The augmented dataset was used to train the SVM, results were tested using only the structurally resolved dataset.

Data preparation by clustering | 3.3.3

Ungapped, heptad-specific multiple alignments of (a) the pool of dimeric and (b) the pool of trimeric PDB samples were performed. Each pool was then divided into clusters such that the maximum sequence identity between any two sequences from two different clusters was 60%. Subsequently, an augmented dataset was created by adding each

sequence from the approved BLAST pool to the cluster of the query sequence from which it originated. Thus, two 60%-clustered datasets were obtained: one based exclusively on PDB samples and one augmented by BLAST.

An identity threshold of 60% was chosen because any lower level would have merged about half of the dataset into a single cluster. This is due to the fact that coiled coils have highly similar secondary structures and thus also have *a priori* a high level of sequence similarity.

Heptad-specific single amino acid frequencies | 3.3.4

For the clustered dataset, each cluster was considered as a single coiled coil sequence. This was accomplished by performing an ungapped, heptad-specific multiple alignment of all sequences in the cluster. Then, a cluster sequence was represented by the relative frequencies of amino acids at each of the aligned positions (analogous to the way clusters are treated when computing BLOSUM matrices [Henikoff *et al.*, 1992]). Finally, the overall single amino acid frequencies were computed as the sums of relative amino acid frequencies at all heptad positions in all clusters.

The statistical significance of each single amino acid position was determined by Fisher's exact test [Fischer *et al.*, 1922], comparing the numbers of occurrences of a given amino acid at a given heptad position in trimers and dimers against the occurrences of other residues in the same heptad position. The total numbers of occurrences of heptad positions in the sequences of the 60%-clustered dataset are approximately 210 in trimers and 800 in dimers. These sample sizes are large enough to have sufficient statistical power to detect even small differences in amino acid frequencies. Finally, 9 single amino acid patterns emerged that were significant according to Benjamini-Hochberg FDR correction [Benjamini *et al.*, 1995] with an FDR threshold of 0.05.

Statistical significance of amino acid pairs | 3.3.5

In order to test whether patterns of pairs of amino acids provide a gain of information compared to single amino acid patterns, all possible pairings of amino acids at specific heptad positions were considered with at most 6 other residues in between. Again, Fisher's exact test was applied, this time comparing joint occurrences of two residues and occurrences of the first residue with other residues (again separately for trimers and dimers). Here, the overall sample size is the number of occurrences of the first single amino acid pattern. Extensive power calculations showed that at least 60 occurrences of a single amino acid pattern are needed to detect statistical differences with sufficient certainty. Of the 4360 pair patterns fulfilling this criterion in the 60%-clustered dataset, 130 pairs showed a p-value of at most 0.05. After applying Benjamini-Hochberg FDR correction and a stringent FDR threshold of 0.05, two pair patterns remained significant. Thus, a gain in information with statistical significance can indeed be observed. Note that many potentially valuable pair patterns may have been overlooked because the sample sizes were too small to detect a difference with sufficient significance.

SVM and kernels | 3.3.6

The non-technical reader may find these introductory tutorials [Burges, 1977; Müller *et al.*, 2001] or standard literature [Vapnik, 1998; Cortes *et al.*, 1986; Christianini *et al.*, 2000] helpful to become familiar with the topic. The functions and algorithms were developed in collaboration with Ulrich Bodenhofer, Ingrid Abfalter, and Sepp Hochreiter (Institute of Bioinformatics, JKU, Linz).

SVM discriminant function | 3.3.6.1

Suppose one wishes to perform a binary classification of samples x_i (in this case amino acid sequences). Each sample can belong either to the positive class with the label $y_i = 1$ (trimers) or to the negative class with the label $y_i = -1$ (dimers). For a given training set with $\{(x_i,y_i) \mid 1 \leq i \leq l\}$, the discriminant function (i.e., the classifier) of the support vector machine is given by

$$f(x) = b + \sum_{i=1}^{l} \alpha_i \times y_i \times k(x, x_i) \quad ,$$

where b is the offset, α_i are the Lagrange multipliers and $k(x,x_i)$ is the kernel.

Coiled coil kernel | 3.3.6.2

$$k(x, y) = \sum_{p} N(p, x) \times N(p, y)$$

A pair pattern p consists of two amino acids and a fixed number of up to m arbitrary amino acids in between. It is indicated at which heptad position the first amino acid must occur: the pattern S.I*f*, for instance, matches a coiled coil sequence if a Ser occurs at an *f* position and an Ile at the next *a* position (with an arbitrary amino acid at the *g* position in between). For a given pattern p and a sequence x, $N(p,x)$ denotes the number of occurrences (i.e., matches) of pattern p in sequence x. The coiled coil kernel calculates the number of coiled coil patterns shared by two sequences, taking multiple occurrences into account. It bears some resemblance to the spatial sample kernel [Kuksa *et al.*, 2008] and the kernel described in [Fong *et al.*, 2004]. However, in contrast to the former, the coiled coil kernel has an additional position/heptad-specific property, and in contrast to the latter, it considers pairs of residues from the same chain and is not restricted to a small set of pairs of

positions. The kernel values were normalized to correct for variations in sequence length [Bodenhofer *et al.*, 2009].

Model selection | 3.3.6.3

The classification performance depends heavily on the choice of model parameters (the coiled coil kernel parameter *m*, the SVM's penalty parameter *C*, raw kernel vs. normalized kernel, unaugmented training set vs. BLAST-augmented training set) and SVM implementations. A common approach is to use cross-validation to select the best set of parameters. This strategy was used employing two well established SVM implementations, the C-SVM implementation of LIBSVM [Chang *et al.*, 2001] and the PSVM [Hochreiter *et al.*, 2006]. The classification results were ranked according to accuracy to guarantee a low misclassification rate. The LIBSVM optimizes margin errors and hence delivers excellent classification at the boundary between the two classes. Thus, it not only provides good classification of new sequences, but is also suitable for classifying and characterizing variations of known sequences that are produced in the course of mutation analysis, as the GCN4 examples described in this work confirm. The PSVM minimizes the mean squared error, which allows balancing the data set by increasing the weights of the labels of the smaller class. This leads, on one hand, to improved separation of whole clusters, but, on the other hand, to more margin errors.

The validity of the model selection procedure was verified by nested cross-validation. In the outer cross-validation loop, the whole dataset was split into 10 parts with a maximum sequence identity of 60% between parts. In each of 10 runs, the ten parts were grouped differently to form a training dataset (9 parts) and an unseen dataset (1 part). Model selection was performed by means of 9-fold inner cross-validation on the training dataset. The resulting best model was then tested on the unseen data. The average test accuracy of the 10 (outer) runs using LIBSVM was 86.9%, which shows that the model selection procedure used yields excellent performance on independent test sets. Hence, cross-validation-based model selection can be safely applied to the

entire dataset. The best model in terms of accuracy obtained in this way resulted from LIBSVM trained with the BLAST-augmented dataset using the normalized coiled coil kernel with $m = 7$ and the SVM penalty parameter $C = 8$. After retraining with the complete dataset (i.e., with no data omitted), this became our PrOCoil model.

Pattern Extraction | 3.3.6.4

Pattern extraction was performed by rearranging the discriminant function $f(x)$ as described in [Bodenhofer *et al.*, 2009] to obtain the weights $w(p)$ of the patterns p, given a support vector machine. y_i denotes the class label ($+1/-1$).

$$f(x) = b + \frac{1}{\sqrt{\sum_p N(p,x)^2}} \sum_p N(p,x) \times \underbrace{\sum_{i=1}^{l} \frac{\alpha_i \times y_i \times N(p,x_i)}{\sqrt{\sum_p N(p,x_i)^2}}}_{=w(p)}$$

Sequence Profiling | 3.3.6.5

The discriminant function $f(x)$ was reformulated such that each position or amino acid i in the sequence x is attributed the weight s_i (i.e., the sum over half of the weight of all patterns of which it is part) it contributes to the discriminant function. The base line of the resulting sequence profiling plot is given by $y = -b/L$.

$$f(x) = b + \sum_{i=1}^{L} s_i = \sum_{i=1}^{L} (s_i - (-\tfrac{b}{L}))$$

Results and Discussion

Cross-reactivity of detection systems | 4.1

The rationale of this approach was to understand the potential cross-reactivity of three common detection systems – 5-(and 6)-carboxy-tetramethylrhodamine (TAMRA), fluoresceinisothio-cyanate in the form of the peptide-bound fluorescein-substituted thiourea derivative (FITC), biotin and streptavidin-POD – with cellulose membrane-bound peptides at the amino acid level. To investigate the potential interaction of these detection systems with individual amino acids, 20 peptides of the sequence GGG$[B]_5$GGG were designed. Herein, $[B]_5$ denotes five repeats of one of the 20 amino acids (for a schematic overview see **Figure 6**). Glycine was used to create non-reactive regions flanking the functional core at the N- and C-termini. This approach generates peptides of reasonable length for the homogeneous display of the defined cores. The peptides were prepared via SPOT synthesis, with each GGG$[B]_5$GGG sequence repeated three times in columns on the peptide array. Additionally, the core motif lengths were varied from $[B]_5$ to $[B]_1$, and also the peptide-specific density was varied in order to identify effects on interaction.

All membrane-bound peptides were analyzed by reversed phase HPLC and MALDI-TOF (see appendix **Table A1**). All masses except those of cysteine-containing peptides were found, and the purity of the SPOT-synthesized peptides was determined (by HPLC) to be in the range of

25% to 85%, which is adequate for screening assays [Kramer *et al.*, 1999; Wenschuh *et al.*, 2000]. Since the masses of the cysteine-containing peptides were incorrect, the analytical results of these peptides were not taken into account.

As soluble interaction partners, peptides of the sequence Gly-Gly-Gly were synthesized, N-terminally modified with biotin, TAMRA, or FITC (label-GGG), and finally purified by HPLC. This tripeptide was used to better meet the assay conditions, because the aforementioned labels are usually chemically coupled to an analyte or a detection antibody.

Peptide arrays containing the core-motifs were incubated *in situ* with a label-GGG and evaluated using optical, fluorescent, and chemiluminescent methods. Strict conditions including short incubation periods and long-time washing procedures were applied to ensure stringency of binding. Binding experiments resulted in measurable spot signal intensities signifying directly or indirectly captured label conjugate.

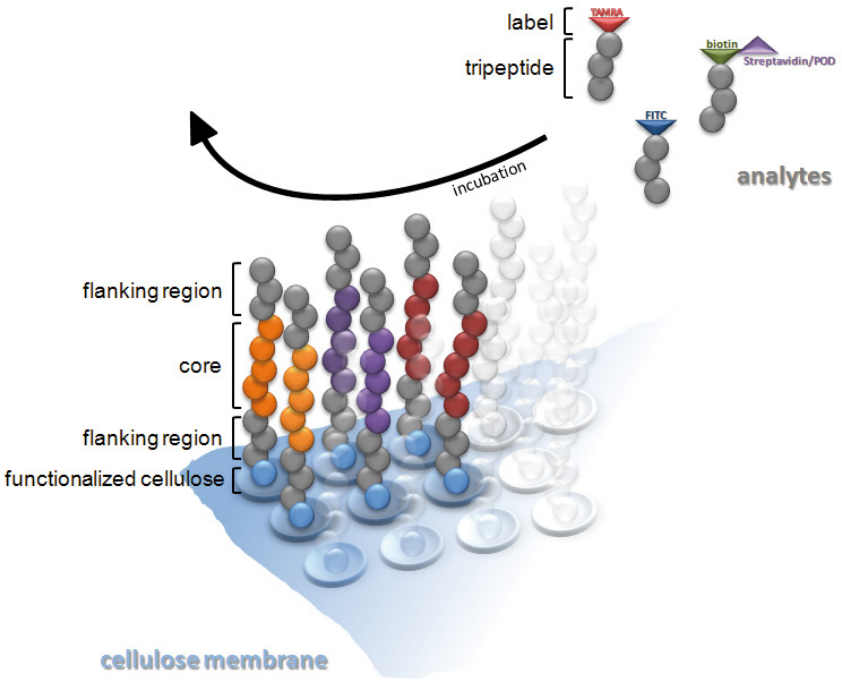


Figure 6 | Schematic peptide and analyte composition. Non-reactive Gly-repeats (gray) flank repeats of core amino acids (colored). Analytes are also composed of non-reactive Gly-repeats labeled with TAMRA (red), FITC (blue), or Biotin (green).

Biotin and streptavidin-POD | 4.1.1

Biotin-labeled samples were used to challenge the peptide arrays and were subsequently detected via streptavidin-POD conjugate using chemiluminescence. Streptavidin-POD was also tested directly on the membrane-bound peptides to differentiate between streptavidin-POD and biotin interactions. **Figure 7** shows that streptavidin-POD is prone to cross-reaction with the positively charged amino acids lysine and arginine.

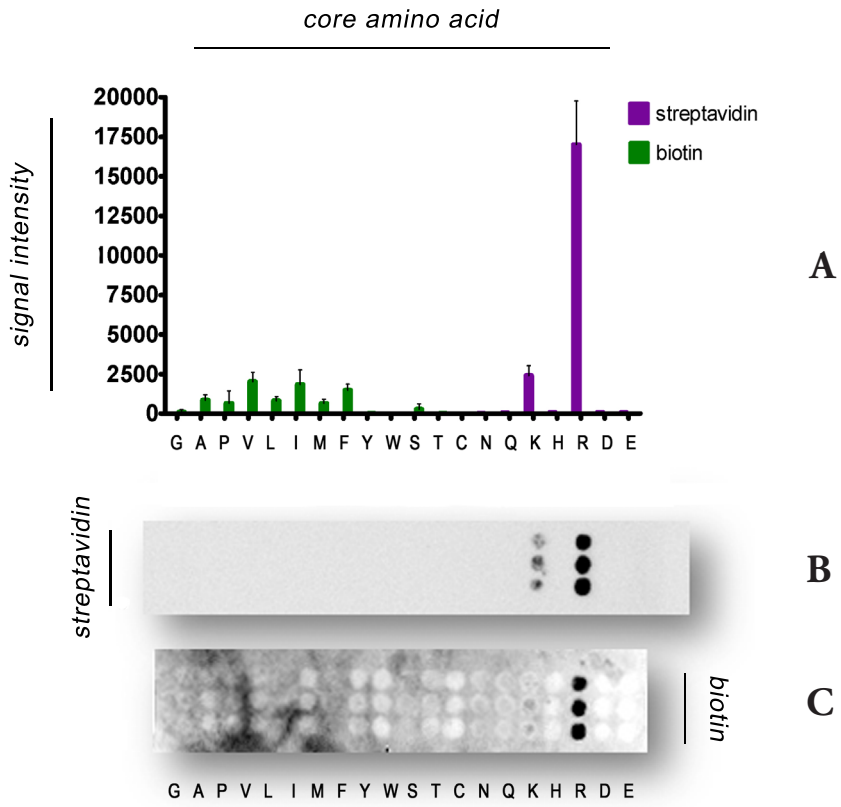


Figure 7 | Streptavidin-POD (purple) and biotin (green) cross-reaction with membrane-bound peptides. (A) The spot signal measured by means of chemiluminescence is calculated from a circular region around the spot center detected in the image. All signals below an SI of 1000 are at the background level and should therefore not be considered interactions between the core and the detection system. Streptavidin-POD results were set as background for the calculation of biotin interactions. Due to the direct interaction of streptavidin-POD with positively charged peptides, any further information about the cross-reactivity of biotin with Lys and Arg has been lost. **(B)** Each spot represents a cellulose membrane-bound peptide of the sequence GGG[B]$_5$GGG, where [B]$_5$ denotes five repeats of one of the 20 amino acids. Black spots denote interactions with streptavidin-POD and **(C)** with biotin-GGG/ streptavidin-POD. The negative control without analyte shows no signal. Error bars represent the standard deviation of three spots.

However, the observed binding is most likely related to streptavidin, as it has previously been shown that peroxidase does not cross-react [Beutling *et al.*, 2008]. Overall, this set of interactions reveals a weak cross-reactive potential of biotin. The bulky aliphatic amino acids valine and isoleucine, and the aromatic amino acid phenylalanine show signals slightly above the background. The smaller amino acids alanine, serine, proline, and leucine show insignificant signals, scarcely visible against the background. Hence, it is reasonable to assume that the small biotin molecule cannot bind to a peptide probe when deeply buried inside the

complex with streptavidin. According to the law of mass action, the interaction between biotin and streptavidin is favored (K_D ~10–15), as the interaction of biotin with the peptides is not supposed to be covalent.

Membrane autofluorescence | 4.1.2

For fluorescent dye-labeled probes it was necessary to consider any spot autofluorescence (membrane and/or peptides). The fluorescence emission of unchallenged peptide arrays in blocking buffer was measured prior to incubation at wavelengths corresponding to label emission (645 nm and 520 nm). As expected, no background signals were detected at 645 nm (see **Figure 10**). The results of the fluorescence spectroscopy (**Figure 8**) reveal the excitation spectra for the full set of peptides with the 20 different cores at 520 nm emission wavelength.

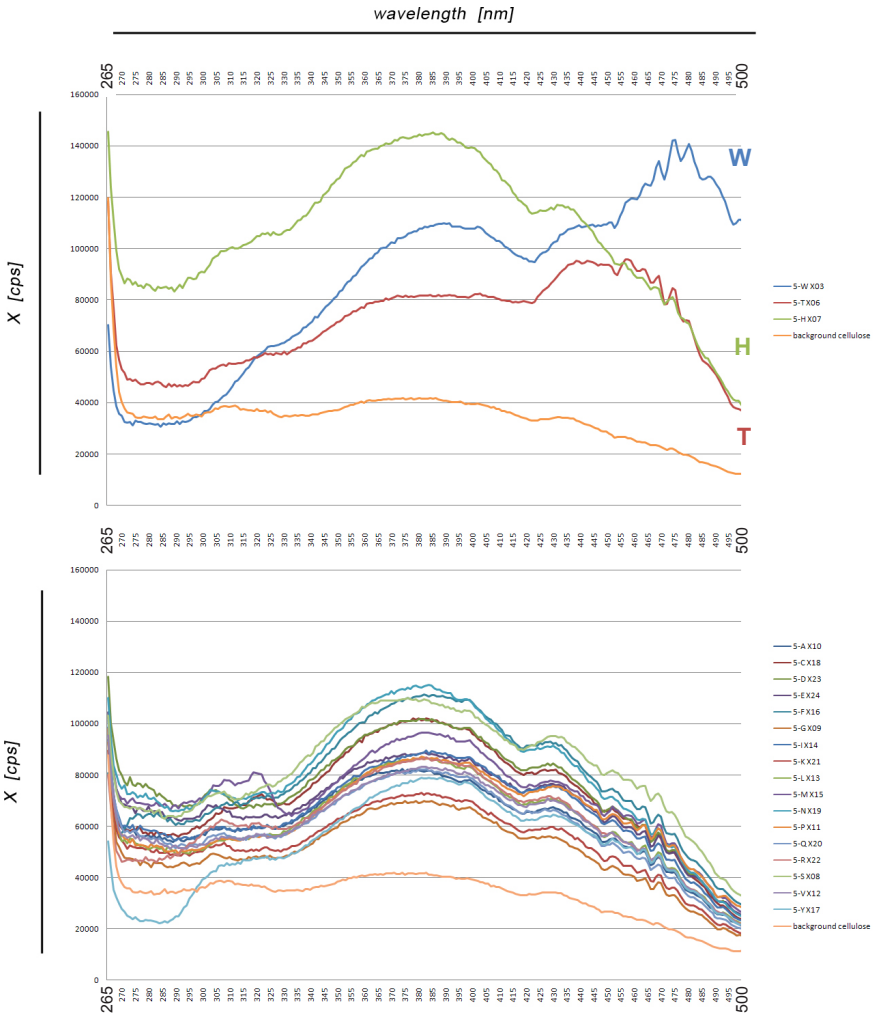


Figure 8 | Spot autofluorescence excitation spectra measured at 520 nm emission wavelength. Spectra of 20 cellulose membrane-bound peptides of the sequence GGG[B]$_5$GGG. The cellulose spectrum (orange) is plotted twice for reference. Excitation spectra (l_x) were measured with the excitation wavelength varying from 260 nm to 500 nm and the emission wavelength (l_m) set to 520 nm. Intensity of fluorescence is expressed as counts/sec.

As shown in **Figures 8, 9A**, and **9B**, common autofluorescence of the cellulose membrane can be observed at 520 nm. This is in accordance with the literature [Reinecke *et al.*, 2005] and may result from membrane impurities that accumulate during the processes of the synthesis cycles, for instance, Fmoc deprotection, side-chain deprotection, or coupling procedures. Additionally, significant spot autofluorescence was measured at 520 nm for Trp-, His- and, unexpectedly, also for Thr-containing peptides (**Figure 9**).

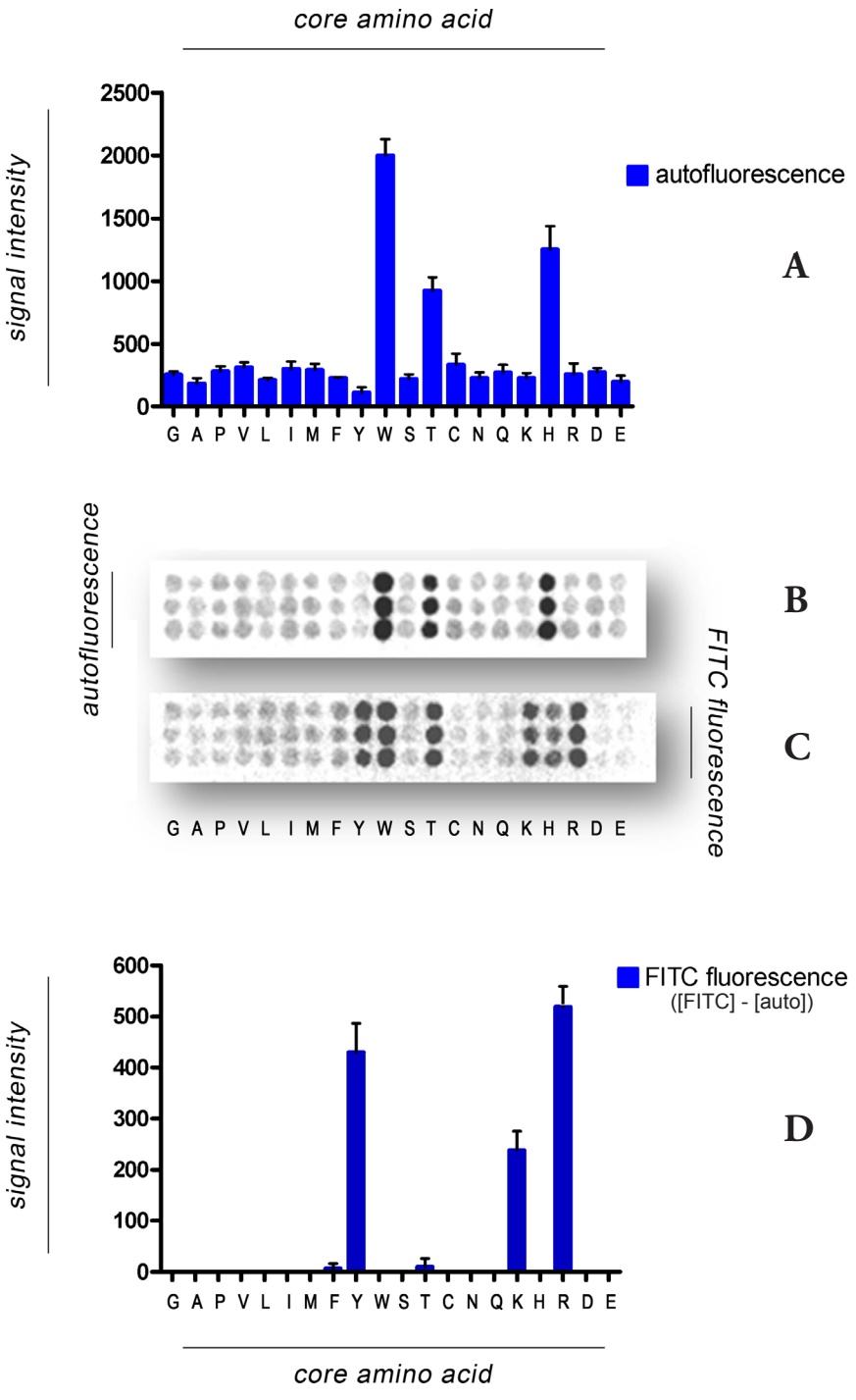


Figure 9 | FITC cross-reaction. (A) Spot autofluorescence at 520 nm. The spot signal measured at 520 nm is calculated from a circular region around the spot center detected in the image shown in B. **(B)** Each spot represents a cellulose membrane-bound peptide of the sequence GGG[B]$_5$GGG, where [B]$_5$ denotes five repeats of one of the 20 amino acids. **(C)** The same array incubated with FITC-GGG, black spots denote interaction, except for the Trp-, Thr-, and Arg-containing spots. **(D)** The spot signal measured at 520 nm is calculated from a circular region around the spot center detected in the image. The signal intensities are background-subtracted (spot autofluorescence at 520 nm). Error bars represent the standard deviation of three spots.

It becomes apparent that the specific spectrum of a peptide closely resembles the spectrum of the cellulose used. It follows that the emission of the cellulose is merely quenched differently by the various peptides. These quenching effects are responsible for the peptide autofluorescence observed and presented in **Figure 9**. All spectra look similar except those of Trp, His, and Thr. While the fluorescence of tryptophan and histidine can be explained by their aromatic ring systems containing more than six valence electrons, the fluorescence observed for the threonine core peptide remains a challenge for interpretation.

FITC | 4.1.3

Probing the peptide array with labeled GGG-peptides and comparing the recorded images with fluorescence records from the unchallenged arrays leads to additional signals. These signals are label-specific and indicate, in the context of this work, sorptive effects of the amino acid core composition. In addition to the background effects mentioned above, arrays challenged with FITC-GGG samples resulted in spot signals at 520 nm for Tyr-, Trp-, Thr-, Lys-, His-, and Arg-containing peptides (**Figure 9C**). After background-correction for membrane autofluorescence, significant signal intensities remained for peptides containing Tyr, Lys, and Arg (**Figure 9D**). These amino acids are therefore interpreted as FITC cross-reactive moieties. Due to the background correction of fluorescence signals at 520 nm, any further information about the cross-reactivity of Trp, His, and Thr was lost.

5(6)-TAMRA | 4.1.4

The results draw a clear picture of the cross-reactivity of amino acid cores with the peptide TAMRA-GGG. As shown in **Figure 10**, significant spot signal intensities at 645 nm were observed for Phe, Tyr, and Trp cores.

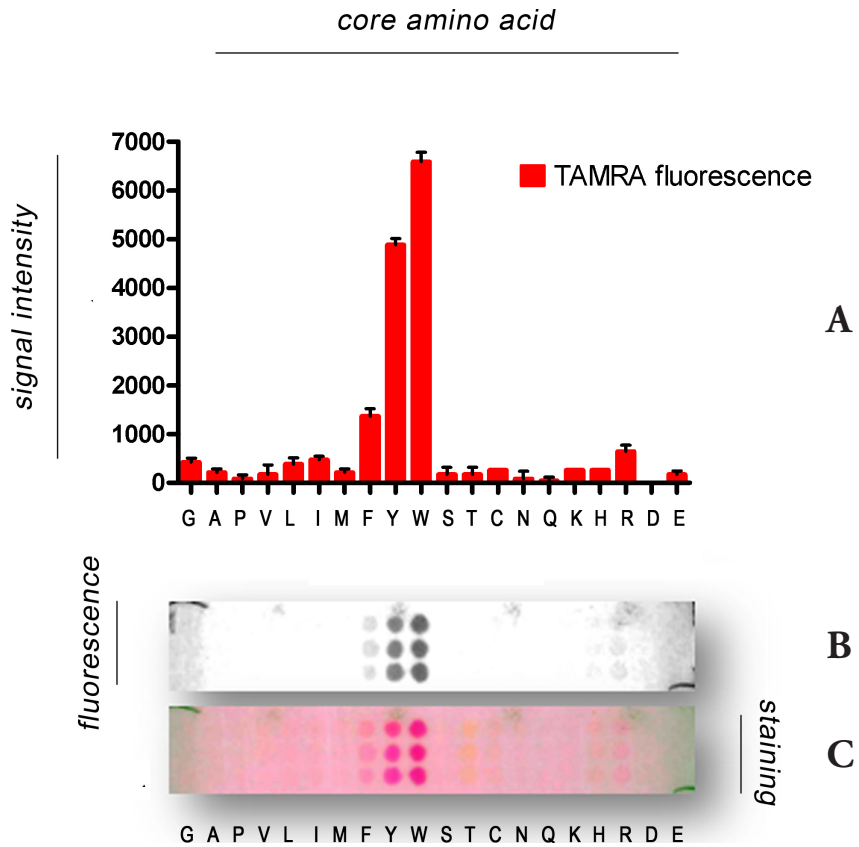


Figure 10 | TAMRA cross-reaction. (A) Fluorescence emission of each corresponding spot measured at 645 nm is calculated from a circular region around the spot center detected in the image. All signals below an SI of 500 are at the background level and should therefore not be considered interactions between the amino acid core and the detection system. **(B)** Fluorescent and **(C)** densiometric read-out. Each spot represents a cellulose membrane-bound peptide of the sequence $GGG[B]_5GGG$, where $[B]_5$ denotes five repeats of one of the 20 amino acids. Contrast was adjusted to ensure better visibility. The negative control without analyte shows no signal. Error bars represent the standard deviation of three spots.

The strength of cross-reactivity between these amino acids and TAMRA follows the order Phe < Tyr < Trp. Additionally, densitometry was used to read the capturing of TAMRA-GGG via staining (see Methods). As shown in **Figure 10B** and **10C**, the results are in accordance with the fluorescence read-out approach. The aromatic TAMRA moiety interacts exclusively with aromatic amino acid cores (**Figure 10A**). Therefore, aromatic stacking is most likely the common driving force for the interaction between amino acid and TAMRA. Stacking is a widespread mechanism for stabilizing organic moieties. It is accomplished by the favorable interaction of π-electrons of aromatic systems [Sygula *et al.*, 2007]. In this case, the π-electron systems of TAMRA and the side group of Trp may interact in an energetically favorable manner via stacking interactions, which the smaller aromatic systems of Tyr and Phe possibly cannot provide to the same extent.

Influential factors | 4.1.5

Peptide-specific density | 4.1.5.1

One open question is the influence of the peptide-specific density per spot on the binding of the label-GGG analytes. Therefore, peptide arrays with various concentrations of GGG[B]$_5$GGG were probed for binding with TAMRA-GGG, FITC-GGG, and Biotin-GGG. The amount of peptide per spot was adjusted as described in the literature [Kramer *et al.*, 1999]. As an example, results for TAMRA are shown in **Figure 11**. The comparable results for FITC and Biotin/Streptavidin-POD are presented in the appendix of this work (**Figures A1** and **A2**).

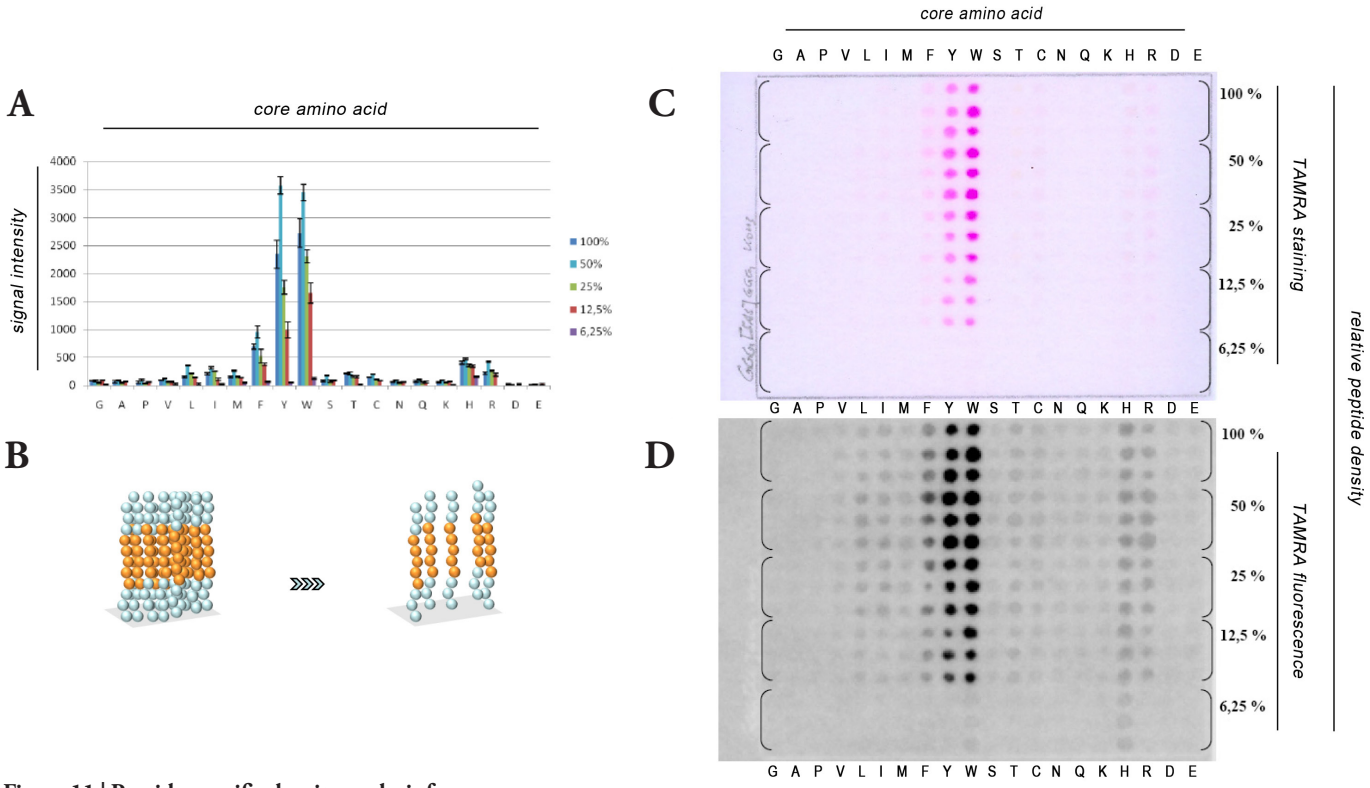


Figure 11 | Peptide-specific density analysis for TAMRA. Concentration library incubated with TAMRA-GGG. **(A)** The spot signal measured at 645 nm is calculated from a circular region around the spot center detected in the image. SI is the calculated mean of three spots. Error bars represent the standard deviation of three spots. **(B)** Schematic overview. **(C)** Densiometric analysis and **(D)** fluorescence at 645 nm. Each spot represents a cellulose membrane-bound peptide of the sequence GGG[B]$_5$GGG, where [B]$_5$ denotes five repeats of one of the 20 amino acids and is repeated three times in different concentrations (100%, 50%, 25%, 12,5%, and 6,25%). Contrast was adjusted to ensure better visibility of the spots.

All detection methods show the reported behavior down to 6.25% of the initial concentration, where the signal breaks off due to the spot's low peptide density. The results suggest that the peptide-specific density influences the signal level whilst not being the cause of the interaction. Reducing the peptide density by a factor of 10 diminishes unwanted side effects. However, it may also result in general binder signal loss. An overall reduction of the peptide load of a membrane is therefore not advisable and must be adapted to the object of research.

Core length | 4.1.5.2

Probing the core reduction peptide arrays in which the core motif lengths vary from $[B]_5$ to $[B]_1$ reveals information about the critical length of the cross-reacting motif. As described above, the core reduction arrays were incubated with TAMRA-GGG, FITC-GGG, and Biotin-GGG. As an example, results for TAMRA are shown in **Figure 12**. The comparable results for FITC and Biotin/Streptavidin-POD are presented in the appendix of this work (**Figures A3** and **A4**).

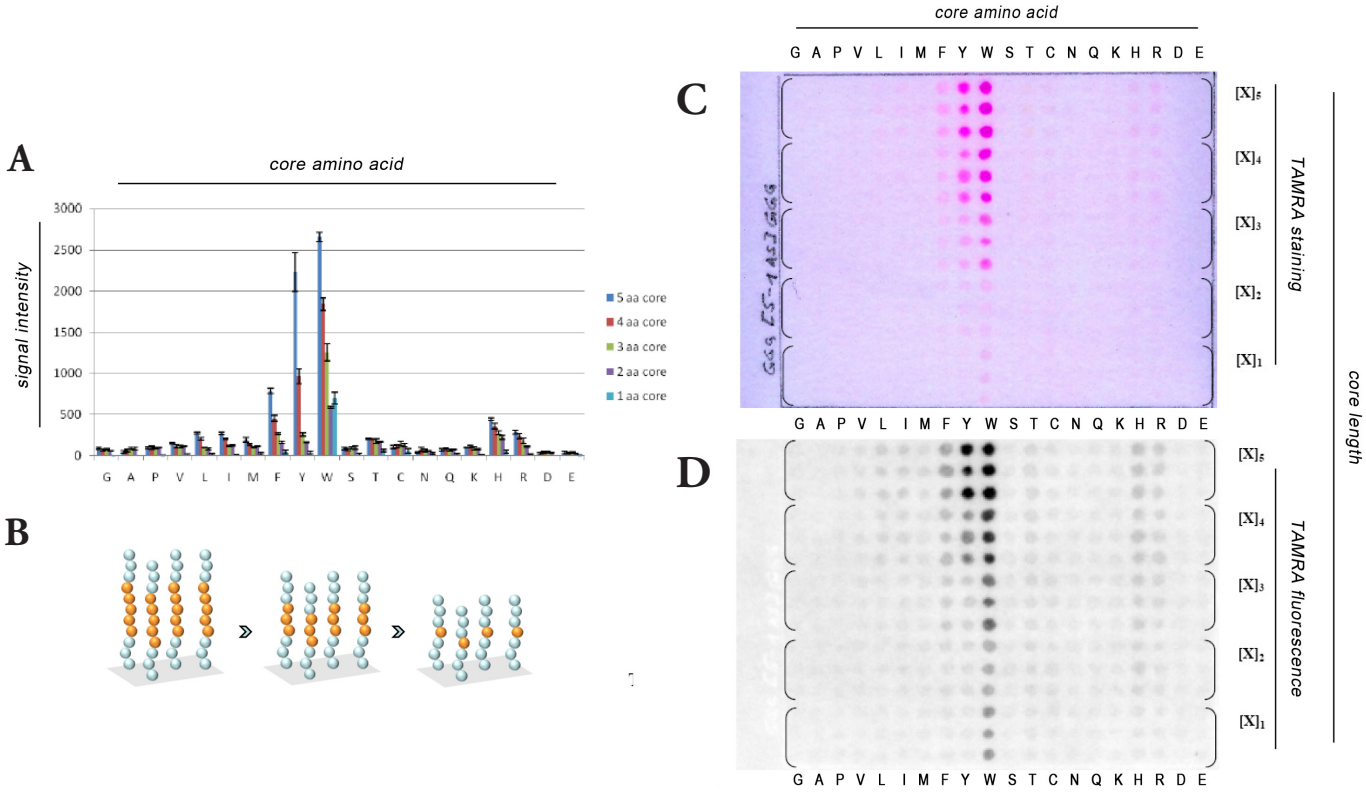


Figure 12 | Core length analysis for TAMRA. Core reduction library incubated with TAMRA-GGG. **(A)** The spot signal measured at 645 nm is calculated from a circular region around the spot center detected in the image. SI is the calculated mean of three spots. Error bars represent the standard deviation of three spots. **(B)** Schematic overview. **(C)** Densiometric analysis and **(D)** fluorescence at 645 nm. Each spot represents a cellulose membrane-bound peptide of the sequence GGG$[B]_5$GGG to GGG$[B]_1$GGG, where $[B]_{5-1}$ denotes 5–1 repeats of one of the 20 amino acids. Every spot is repeated three times. Contrast was adjusted to ensure better visibility of the spots.

A strong dependency on the quantity of aromatic amino acids can be observed for TAMRA. In the case of Trp, signals can be detected even when the core is reduced to one amino acid. The reduction library incubated with FITC-GGG (**Figure A3**) shows signals above the background for all cross-reacting core reductions, the intensities of which decrease with the length of the core. The biotin/streptavidin-POD analysis (**Figure A4**) reveals that interactions with Val, Leu, Ile, and Phe occur only if the cross-reacting amino acid is repeated more than four times. Interaction with positively charged amino acids remains observable at the critical length of two (Lys) or even just one (Arg) core position.

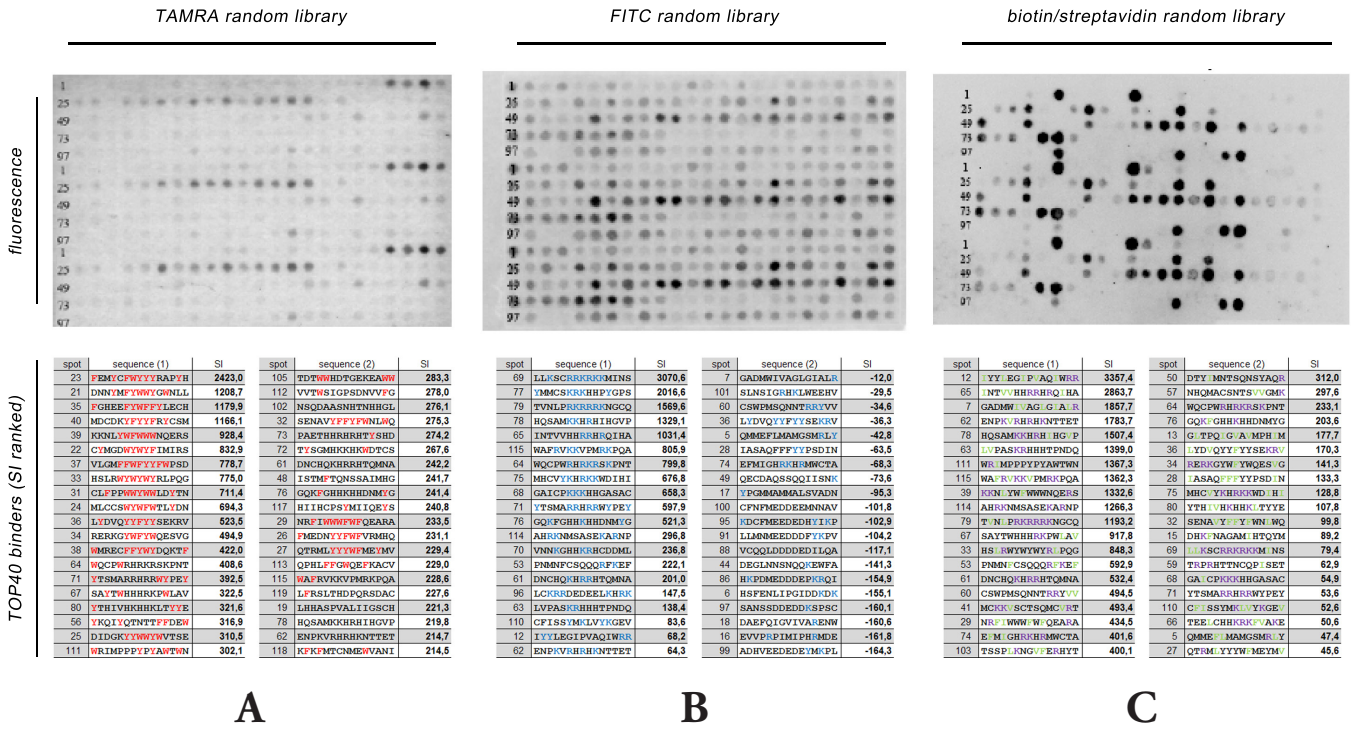


Figure 13 | Random libraries for TAMRA, FITC, biotin/streptavidin-POD. Comparison of random peptide libraries incubated with different detection systems. All arrays (spots 1 to 120) are repeated three times resulting in three identical subarrays. **(A) Top**: Random peptide library incubated with TAMRA-GGG. Each of the 120 spots represents a cellulose membrane-bound 15-meric peptide of random sequence with weighted cores. Contrast is adjusted to ensure better visibility of the spots. **(A) Bottom**: Top 40 sequences (SI sorted). The spot signal measured at 645 nm is calculated from a circular region around the spot center detected in the image. Trp, Tyr, and Phe are highlighted in red. SI is the calculated mean of three spots. **(B) Top**: Random peptide library incubated with FITC-GGG. Contrast was adjusted to ensure better visibility of the spots. **(B) Bottom**: The spot signal measured at 520 nm is calculated from a circular region around the spot center detected in the image. Tyr, Arg, and Lys are highlighted in blue. SI is the background- (i.e., autofluorescence-) corrected calculated mean of three spots. **(C) Top**: Random peptide library incubated with biotin-GGG and streptavidin-POD. Each of the 120 spots represents a cellulose membrane-bound 15-meric peptide of random sequence with weighted cores and is repeated three times. Contrast was adjusted to ensure better visibility of the spots. **(C) Bottom**: The spot signal measured by means of chemiluminescence is calculated from a circular region around the spot center detected in the image. Arg and Lys are highlighted in purple; Val, Leu, Ile, and Phe are highlighted in green. SI is the calculated mean of three spots.

The approach using model peptides leads to conclusive results. However, these findings have to be verified in a more realistic setting. Therefore, a 15-meric random peptide library of 120 peptides with physicochemically weighted cores and random flanking residues was designed.

These sequences were SPOT synthesized in triplication on a cellulose membrane and were probed, freshly prepared, for binding TAMRA-GGG (**Figure 13A**), FITC-GGG (**Figure 13B**), and Biotin-GGG (**Figure 13C**), respectively. Interestingly, in many cases of this set-up just two physicochemically similar cross-reactive amino acids close to each other suffice to observe the above-mentioned effects that were revealed using model peptides. The density and frequency of the cross-reactive amino acids correlate with the intensity of the measured spot signals. A comparison shows that the signal intensity of each of the 120 spots varies significantly depending on the detection method used.

Contribution of the tripeptide-analyte | 4.1.5.4

The tripeptide Gly-Gly-Gly itself does not contribute to the overall interaction of the conjugated construct (label-GGG) with membrane-bound peptides, as no evidence was found for experiment-spanning recurring signals that would indicate binding events of the Gly-Gly-Gly peptide (compare **Figures 7**, **9**, and **10**). Thus, measuring specifically the direct influence of the detection system on the 20 core positions of the membrane-bound peptide probes is possible.

Analysis of coiled coil association | 4.2

Single-substitution analysis | 4.2.1

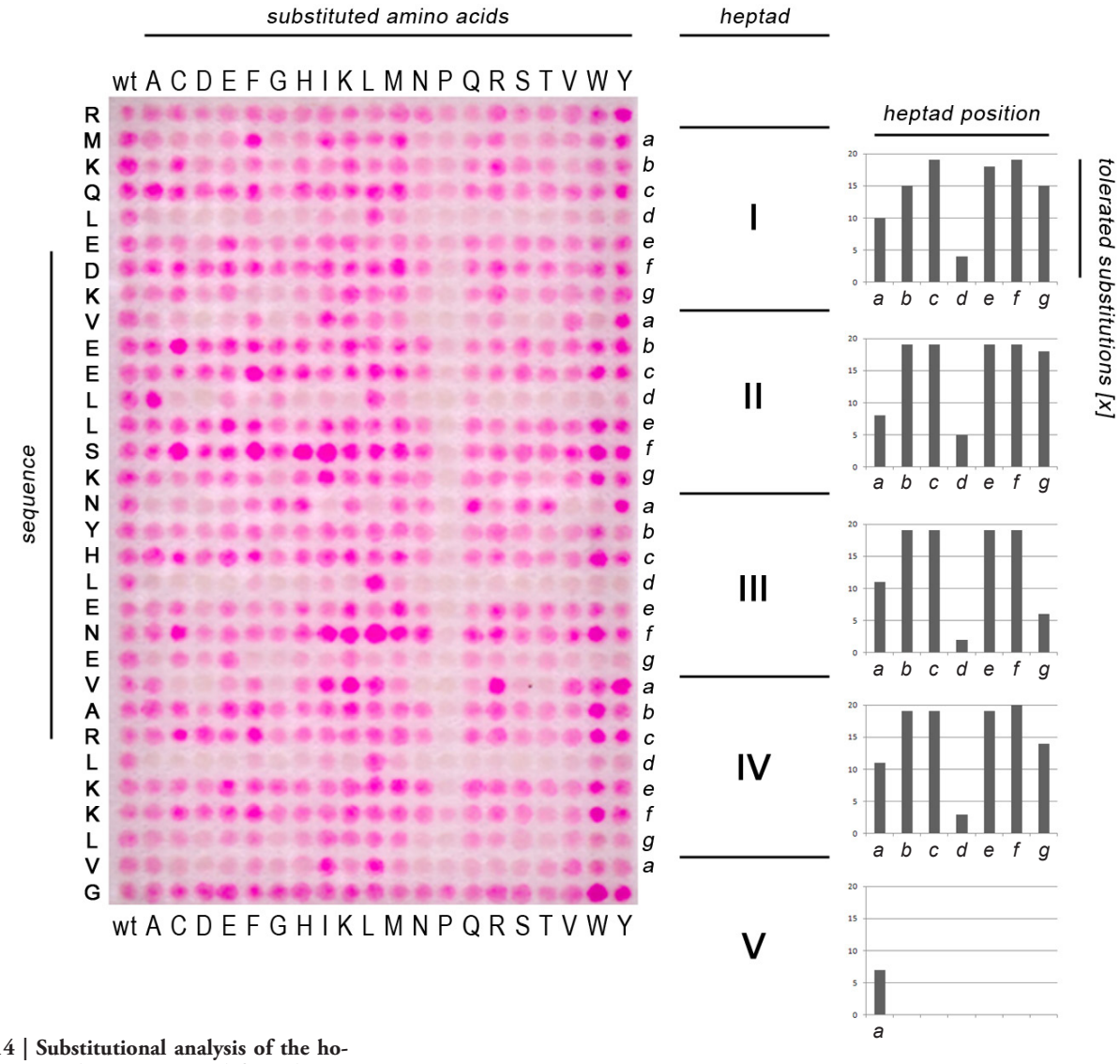


Figure 14 | Substitutional analysis of the homomeric GCN4 leucine zipper. (left) Red spots denote interactions between cellulose membrane-bound variants and a dye-labeled wildtype GCN4 leucine zipper sequence that was synthesized by standard solid-phase peptide synthesis and labeled with TAMRA at the N-terminus. Each spot corresponds to a variant in which one residue of the *wt* sequence given at the top was replaced by one of the 20 gene-encoded amino acids as specified on the left. Spots in the first row represent the *wt* sequence. **(right)** All spot signals of the array shown on the left were measured quantitatively, and successful replacements (countable binding spots) were determined. The quantity of tolerated substitutions is plotted against the positions in a given heptad.

Peptide arrays comprising all single point substitution variants of the GCN4 (**Figure 14**) and c-Fos leucine zippers (**Figure 15**) were probed for binding to wildtype (*wt*) GCN4 and wildtype c-Jun respectively. For visualization, both the $GCN4_{wt}$ and the $c\text{-}Jun_{wt}$ sequences were labeled N-terminally with TAMRA. GCN4, c-Fos, and c-Jun belong to the family of bZip transcription factors, and dimeric coiled-coil folding is essential for their biological function. As shown in **Figure 14**, both homomeric and heteromeric associations are observed for the GCN4 interaction, while only heteromeric associations are found for the c-Fos/c-Jun interaction (**Figure 15**).

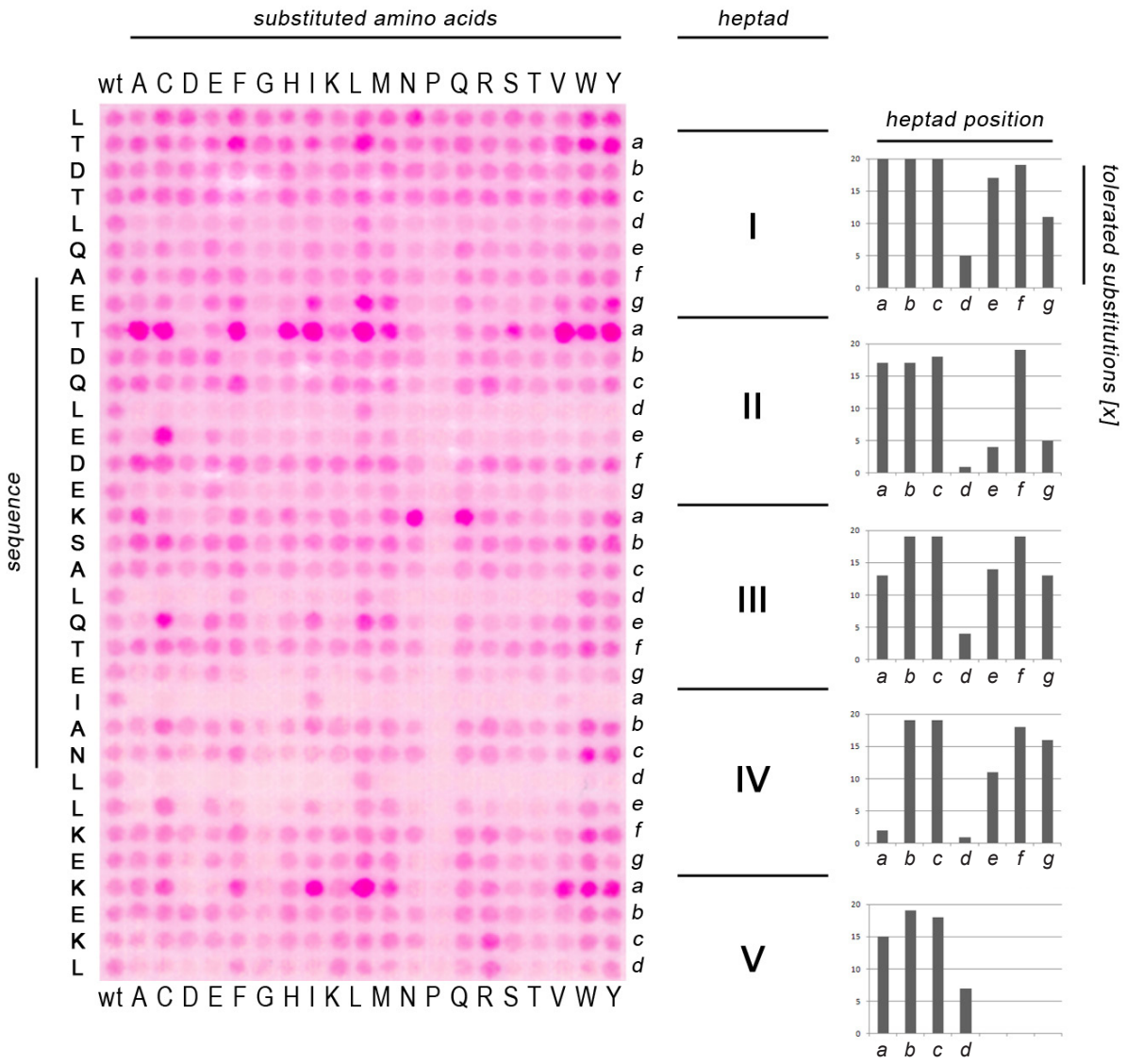


Figure 15 | Substitutional analysis of the heteromeric c-Fos/c-Jun domain. (left) Red spots denote interactions between cellulose membrane-bound variants of c-Fos and a dye-labeled wildtype c-Jun domain that was synthesized by standard solid-phase peptide synthesis and labeled with TAMRA at the N-terminus. Each spot corresponds to a variant in which one residue of the c-Fos$_{wt}$ sequence given at the top was replaced by one of the 20 gene-encoded amino acids as specified on the left. Spots in the first row represent the c-Fos$_{wt}$ sequence. **(right)** All spot signals of the array shown on the left were measured quantitatively, and successful replacements (countable binding spots) were determined. The quantity of tolerated substitutions is plotted against the positions inside a given heptad.

The GCN4 substitution array consists of 651 peptides (589 substitution variants and 2 × 31 *wt*-sequences). The binding experiment resulted in 411 association events, corresponding to 63% of the array population. The 693-peptide c-Fos array (629 substitution variants and 2 × 33 *wt* sequences) resulted in 469 c-Jun associations, which corresponds to 68% of the array population. The helical character of the GCN4 and c-Fos sequences is confirmed by the fact that substitutions with the helix-breaking amino acid Pro are not tolerated within the domain except in N- and C-terminal positions. As expected and depicted in the corresponding variability plots, leucine at core positions d_I–d_{IV} has very low variability, with the exceptions of Leu12 (d_{II}), which can be replaced by Ala in the case of GCN4, and Leu19 (d_{III}), which can be replaced by Trp in the c-Fos/c-Jun dimer. In contrast, core positions *a* are predominantly of intermediate variability, with the exception of

c-Fos Ile23 (a_{IV}), which cannot be replaced. Notably, a significant number of c-Fos *a* substitutions resulted in stronger c-Jun associations compared to the wildtype complex. Substituting the only hydrophilic residue inside the core, asparagine, at position (a_{III}) in GCN4, with one of the hydrophobic branched chain amino acids leucine, isoleucine, or valine resulted in variants showing no association with the wildtype GCN4 sequence. However, different substitutions, for instance, with bulky tyrosine, were tolerated in this position. Interestingly, replacement of the wildtype Lys16 in the comparable c-Fos-(a_{III})-position with asparagine or glutamine resulted in strong c-Jun association. The array analysis suggests that the c-Fos *a* positions play a critical role in stabilizing the heteromeric coiled-coil interaction with c-Jun. In both arrays, the variability of the core-flanking positions *g* and *e* depends on which heptad they are located in. In the case of c-Fos (**Figure 15**), positions *g* and *e* of the second heptad show low variability, whereas *g* in the third heptad of GCN4 was found to be crucial and tolerated no amino acid replacement.

To distinguish coiled coil from non-coiled-coil interaction experimentally, the specific structural characteristics (i.e., the invariable hydrophobic core positions) of coiled coils become relevant. As demonstrated here, the hydrophobic core is a very selective region. Substitution of an amino acid, especially at the very sensitive *d* position, often leads to disruption of the coiled coil. However, exactly this separates coiled coil from non-coiled-coil interaction and provides a means to distinguish coiled-coil from other interactions in a high throughput approach. Further coiled coil substitution analyses of the Jun homodimer (**Figure A5**), of c-Jun with c-Fos (**Figure A6**), and of a designed [Burkhard *et al.*, 2000] 15-meric coiled coil (**Figure A7**) can be found in the appendix of this work. The described behavior can be observed in every case.

Double-substitution analysis | 4.2.2

As multiple substitutions can entail a switch from dimeric to trimeric or tetrameric structures [see, e.g., Harbury *et al.*, 1993], the peptide array approach was extended to investigate double substitutions, focusing on residues within, or in close vicinity of, the core. Double substitutions at positions *a*/*d* and *a*/*e* of the GCN4 leucine zipper sequence resulted in a synthetic peptide array of double-substitution variants, which were probed for binding to the $GCN4_{wt}$ sequence (**Figure 16A**).

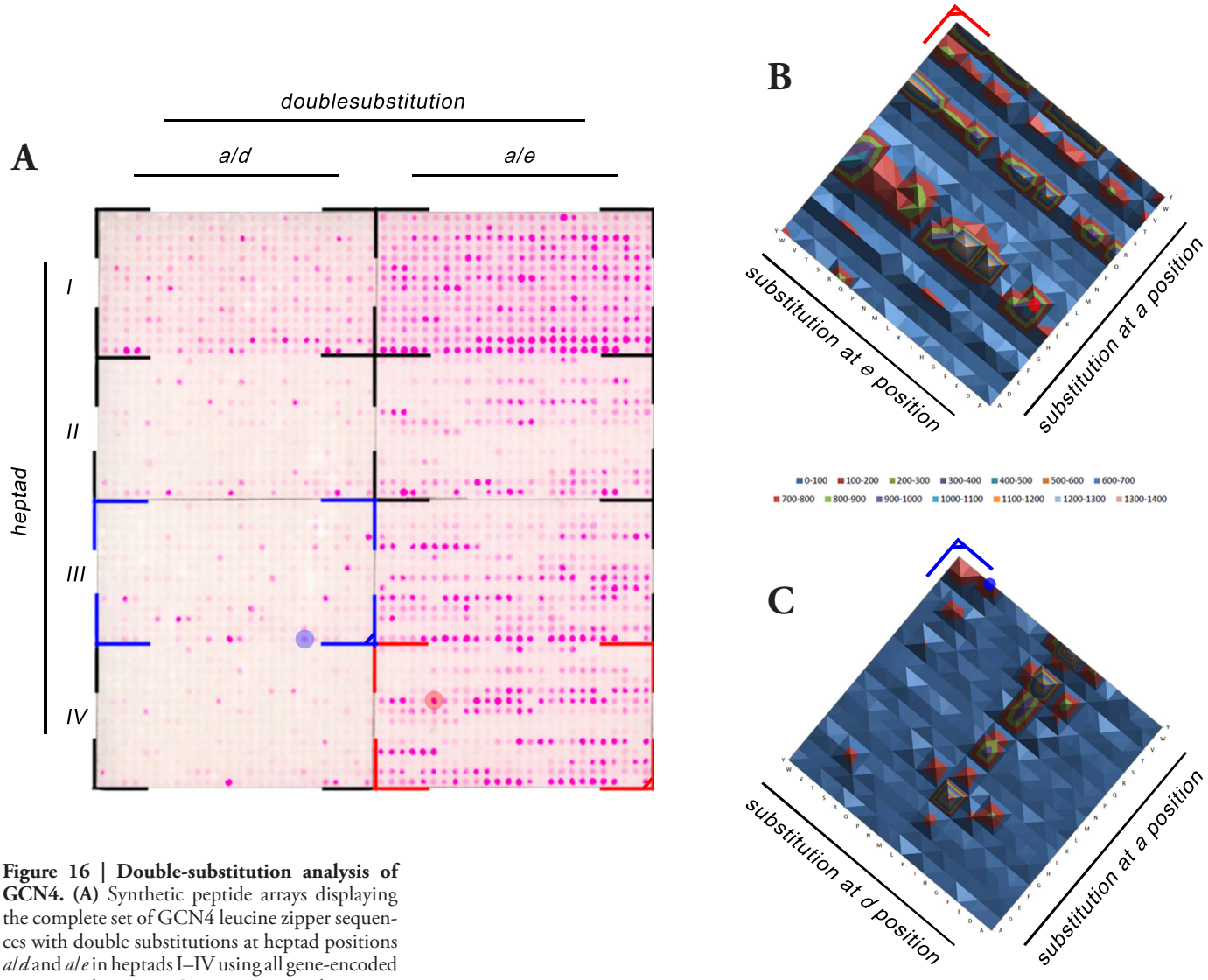


Figure 16 | Double-substitution analysis of GCN4. (A) Synthetic peptide arrays displaying the complete set of GCN4 leucine zipper sequences with double substitutions at heptad positions *a*/*d* and *a*/*e* in heptads I–IV using all gene-encoded amino acids except Cys. For practical reasons, each assembly comprised 26×14 synthesis sites, thus enabling the implementation of *wt* controls. Each colored spot represents a variant that is associated heterospecifically with a wildtype GCN4 leucine zipper domain that is marked with TAMRA at the N-terminus. Heat map diagrams depicting the quantitatively measured SIs for **(B)** *a*/*e* and **(C)** *a*/*d* replacements in heptads IV and III, respectively. The SIs corresponding to the colors are displayed at the bottom of each heat map. The dimeric $GCN4_{V23K,K27E}$ mutant is highlighted in red. The trimeric $GCN4_{N16Y,L19T}$ mutant is highlighted in blue.

Heptad I is characterized by the highest tolerance of substitution, whereas heptads II–IV show different substitution tolerances following the order I > III > II > IV for *a*/*d* and I > III > IV > II for *a*/*e*. Given that double substitutions at *a*/*e* are variable, it is not surprising that they also resulted in the highest number of associations, implying that single-point substitutions at positions *a* and *e* act additively upon

simultaneous exchange. Surprisingly, however, *a*/*d* substitutions were equally poorly tolerated, which confirms the aforementioned results concerning the invariability of the *d* position in coiled coils.

Figure 16B and **16C** depicts the substitution matrix of two example double-substitutions. The heatmap for *a*/*e* substitutions in heptad III (**Figure 16B**) reveals that the *a* position tolerates substitutions with Ile, Lys, Leu, Arg, Val, and Tyr. The *e* position is highly variable in these combinations. However, other combinations are hardly tolerated. The *a*/*d* substitution heatmap for heptad IV (**Figure 16C**) depicts a much more invariable substitution pattern. At *d* position, the wildtype Leu is tolerated almost exclusively, while the partnering *a* position is highly variable. Interestingly, Tyr is one of the very few other amino acids tolerated in the *a* position. Its presence opens up the variability of the *d* position, and in addition to Leu, also Ile, Lys, Ser, The, Val, Trp, and Tyr are tolerated.

As analytical ultracentrifugation revealed (**Table 1** and **Figures 22-24**), several mutations lead to variants that trimerize, for example, the highlighted $GCN4_{N16Y,L19T}$ (blue), while others, such as $GCN4_{V23K,K27E}$ (red), are dimeric like the *wt* sequence.

mutant	K_2 (M^{-1})	K_3 (M^{-2})
$GCN4_{N16Y,L19T}$	936 765/1136	$1.2 \cdot 10^7$ $1.15 \cdot 10^7 / 1.25 \cdot 10^7$
$GCN4_{V23K,K27E}$	182 165/201	n.a.

Table 1 | Homoassociation constants of selected GCN4 mutants. $GCN4_{N16Y,L19T}$ is a trimer and $GCN4_{V23K,K27E}$ is a dimer.

However, in this case, the SPOT technology has reached its limit. Despite being a high throughput method with various applications and possibilities for analyzing coiled coils, it cannot provide information about the oligomeric state of the peptides. A potential approach to solving the problem, i.e., determining the relationship between analyte concentration and the stoichiometry of the investigated coiled coils, was tested (data not shown). This approach was based on the assumption that varying the concentration of the analyte results in changes in signal intensity that correspond to its oligomeric state. However, inconclusive results from the concentration series experiments led to the need for an alternative approach that elucidates the behavior of coiled coils.

Analysis of coiled coil oligomerization | 4.3

As previously demonstrated, synthetic peptide arrays are well suited to studying coiled-coil associations, but they do not provide information about the stoichiometry of the coiled coils. To uncover the factors that influence the oligomerization of coiled coils, biochemical and biophysical approaches must be complemented with bioinformatics methods.

Data preparation | 4.3.1

The PDB was scanned for coiled coil segments, and thus a database of 385 dimeric and 92 trimeric sequences was created. Other oligomers were not considered in this work, since they account for less than 10% of the structurally resolved coiled coils. To augment this set by newly sequenced genome data, a sophisticated BLAST [Altschul *et al.*, 1990] approach with stringent filtering was employed, which resulted in a combined dataset of 2043 dimeric and 791 trimeric sequences. In contrast to hitherto published approaches to coiled coil analysis, the data was clustered to minimize statistical bias. Therefore, conclusions based on the dataset are representative of coiled coil sequences in general and are not biased by the variability of a few sequences that are in the limelight of scientific interest and thus more prevalent in the PDB. Both the PDB and the augmented dataset were used to create clustered datasets with a 60% sequence-identity threshold, i.e., the maximum sequence identity between any two sequences of two different clusters is 60%.

Pattern identification by statistical analysis | 4.3.2

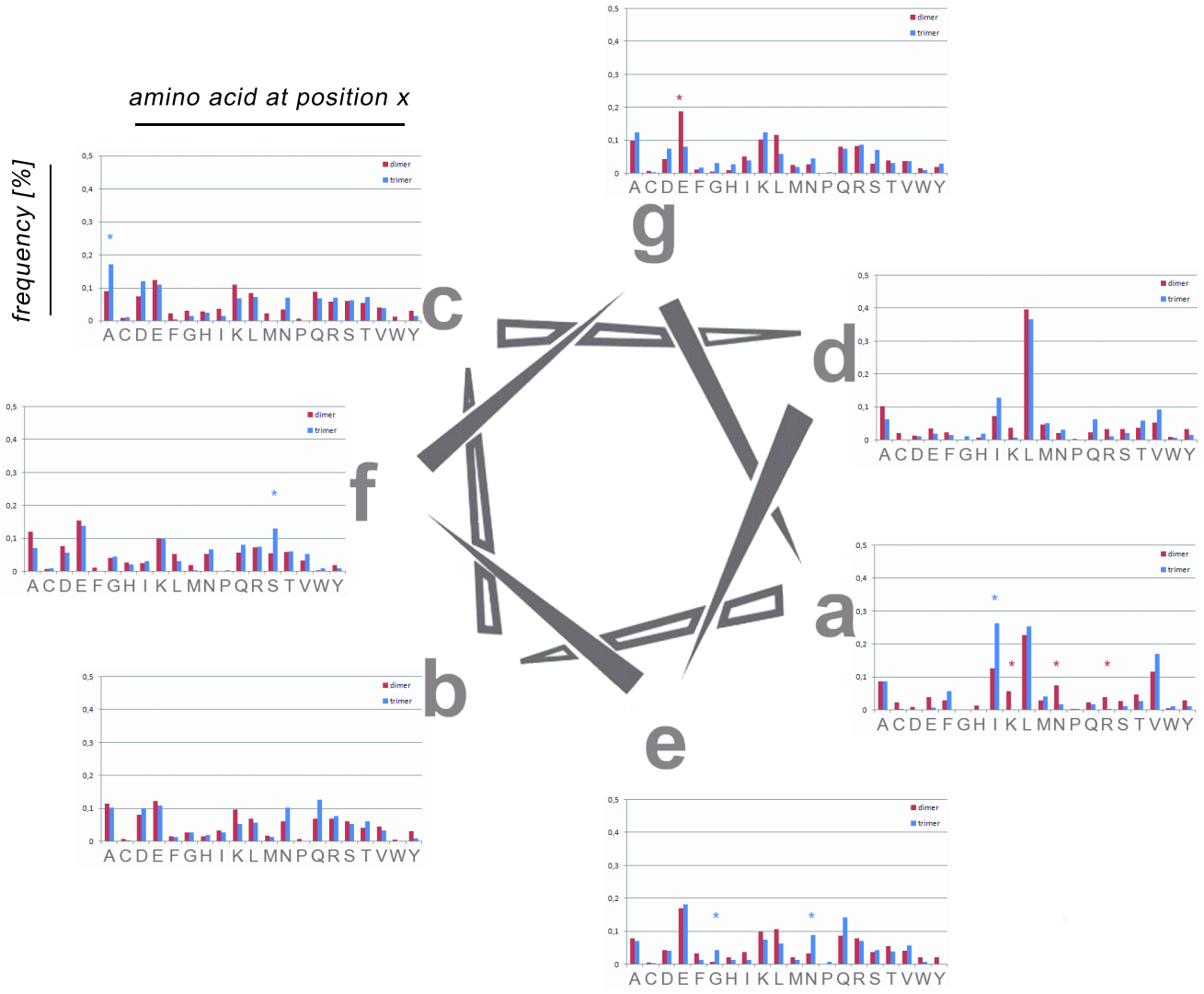


Figure 17 | Relative frequency of each amino acid at a specific heptad position. The amino acid distributions of dimers are displayed in red and those of trimers in blue. To provide a better overview, the charts for each position are assigned to a helical wheel diagram. The frequency of occurrence is plotted against the amino acid at a given position inside a heptad. Significant p-values for dimers (red) and trimers (blue) according to Benjamini-Hochberg correction are highlighted with asterisks.

The first method employed in search of oligomerization rules was a statistical analysis of the frequency of each amino acid at each position of the heptad register in dimers and trimers, in line with Woolfson & Alber's first oligomerization predictor [Woolfson *et al.*, 1995]. The relative frequency results for the 60%-clustered dataset are shown in **Figure 17**. Using Fisher's exact test [Fisher, 1922] p-values were calculated in order to identify those amino acids at specific heptad positions whose comparatively higher frequencies in one oligomer than in the other were statistically significant. Correction for the false discovery rate resulted in 9 significant residues at specific heptad positions according to the Benjamini-Hochberg method [Benjamini *et al.*, 1995] (see **Table 2**).

Initially, it may seem that there is a clear preference for Ile at both hydrophobic core positions of trimers, as previously described in the literature [Harbury *et al.*, 1993]. However, a clear and significant preference for the amino acid Ile only in *a* positions of the hydrophobic cores of trimers could be identified. This β-branched amino acid also frequently occupies the *d* positions in trimers, but false discovery rate correction shows that the high prevalence in *d* positions is, in fact, not statistically significant.

pattern	indicated class	*p*-value	corrected *p*-value (BH)
I*a*	trimer	$3.92 \cdot 10^{-7}$	$5.48 \cdot 10^{-5}$
K*a*	dimer	$5.49 \cdot 10^{-6}$	$3.84 \cdot 10^{-4}$
E*g*	dimer	$1.01 \cdot 10^{-4}$	$4.71 \cdot 10^{-3}$
N*a*	dimer	$2.90 \cdot 10^{-4}$	$1.02 \cdot 10^{-2}$
G*e*	trimer	$5.13 \cdot 10^{-4}$	$1.44 \cdot 10^{-2}$
S*f*	trimer	$5.48 \cdot 10^{-4}$	$1.28 \cdot 10^{-2}$
A*c*	trimer	$1.53 \cdot 10^{-3}$	$3.06 \cdot 10^{-2}$
N*e*	trimer	$1.67 \cdot 10^{-3}$	$2.92 \cdot 10^{-2}$
R*a*	dimer	$3.11 \cdot 10^{-3}$	$4.83 \cdot 10^{-2}$

Table 2 | Statistically significant single amino acid patterns. Computed from the 60%-clustered dataset. FDR correction of p-values according to Benjamini & Hochberg (BH) resulted in 9 significant patterns. All other patterns were above the significance threshold of 0.05. S*f*, for instance, denotes a Ser (S) at an *f* position in a heptad.

As shown in **Table 2**, Arg, Asn, and Lys at *a* positions are the only amino acids in a hydrophobic core position that have a statistically higher prevalence in dimers. Interestingly, these residues in this particular heptad position are tolerated almost exclusively in dimers.

In trimers, positions that usually form salt bridges show a comparatively higher prevalence of the small, uncharged amino acid Gly at *e* position. Additionally, Asn is more prevalent at this position in trimers. In dimers, Glu in *g* position, but not in *e* position, is statistically significant. In the literature [see, e.g., O'Shea *et al.*, 1993], *e* and *g* positions are usually attributed the same characteristics. The results presented in this work, however, show that the amino acid distributions at these heptad positions are, in fact, different. For positions that do not participate in direct coiled coil interaction, statistically significant amino acids could only be found in trimers, namely Ala at *c*, and Ser at *f* positions.

In summary, there are obvious statistical differences in the occurrences of certain residues at certain heptad positions in dimers and trimers. However, these differences are insufficient to distinguish dimers from trimers, as the number of possible combinations of single amino acids

at specific positions is too big and the set of indicators (see **Table 2**) too small. Specific residues at certain positions occur relatively infrequently, giving these occurrences a high specificity but low sensitivity, i.e., for the majority of sequences the oligomerization state cannot be predicted because there is a lack of reliable indicators. Moreover, since trimer indicators may also be found in dimers (see Figure 17), several indicators in combination are required to classify a sequence.

Against this background, approaches based on single amino acid statistics seemed too simple to capture the complexity of oligomerization. Hence, the focus was shifted to the dependencies between amino acids within and beyond a heptad. In order to verify that oligomeric tendency is shifted significantly by considering an additional amino acid at another position, we applied Fisher's exact test to the 60%-clustered dataset. We found that interactions have significant influence on oligomerization, even after correcting for the false discovery rate according to Benjamini-Hochberg (see **Table 3**).

single pattern	pair pattern	indicated class	*p*-value	corrected *p*-value (BH)
E*e*	E..I*e*	trimer	$1.41 \cdot 10^{-6}$	$6.16 \cdot 10^{-3}$
E*b*	E.....Q*b*	trimer	$6.08 \cdot 10^{-6}$	$1.33 \cdot 10^{-2}$
V*a*	V....S*a*	trimer	$4.82 \cdot 10^{-4}$	$7.00 \cdot 10^{-1}$
L*d*	L...N*d*	dimer	$4.89 \cdot 10^{-4}$	$5.33 \cdot 10^{-1}$
E*g*	E...I*g*	trimer	$4.95 \cdot 10^{-4}$	$4.32 \cdot 10^{-1}$
A*a*	A..A*a*	trimer	$5.70 \cdot 10^{-4}$	$4.14 \cdot 10^{-1}$
L*d*	L.....A*d*	trimer	$6.93 \cdot 10^{-4}$	$4.32 \cdot 10^{-1}$
A*g*	AI*g*	trimer	$9.87 \cdot 10^{-4}$	$5.38 \cdot 10^{-1}$
K*e*	K.....T*e*	trimer	$1.07 \cdot 10^{-3}$	$5.17 \cdot 10^{-1}$
L*d*	L...V*d*	trimer	$1.26 \cdot 10^{-3}$	$5.51 \cdot 10^{-1}$

Table 3 | Statistically significant pairwise amino acid patterns. The 10 pair patterns that provide the most significant information gain compared to the single amino acid patterns at their first positions. The statistics were derived from the 60%-clustered dataset. FDR correction according to Benjamini & Hochberg (BH) resulted in 2 statistically significant pairwise amino acid patterns. Note, however, that only those pair patterns were considered for which the number of samples was sufficiently high to detect a significant difference. E..I*e*, for instance, denotes a pattern with Glu (E) at an *e* position, Ile (I) at the next *a* position, and two arbitrary amino acids in between.

Inspired by Berger *et al.*, who used pairwise residue correlations for predicting coiled coils [McDonnell *et al.*, 2006; Berger *et al.*, 1995] and made progress in predicting dimer and trimer formation [Wolf *et al.*, 1997], this approach is based on the aforementioned hypothesis that all amino acids in a given sequence influence each other. Thus, a method that could draw on a maximum number of interactions and combine these into a network of rules would allow predicting and examining oligomerization. Support vector machines (SVMs) are ideally suited to this task and have previously been used in a different context to predict protein-protein interactions that are mediated by the coiled coil motif [Fong *et al.*, 2004].

Model selection and classification results | 4.3.3

To identify the SVM classifier (i.e., model) with the optimum SVM and coiled coil kernel parameters, it had to be assessed which model performs best on future (previously unseen) data. For this purpose, nested 10-fold cross-validation was applied, as is common practice. In 10-fold cross-validation, the data pool (all structurally resolved and 60% clustered PDB samples) is divided into 10 parts. Each part is withheld once to act as an unseen test dataset, while the remaining 9 parts (PDB samples plus their corresponding BLAST samples) are used for selecting the best model. Finally, the SVM and kernel parameters were chosen for which the best classification results in terms of accuracy were achieved. The results in **Table 4** show how the sensitivity-specificity trade-off manifests differently in models resulting from different SVM implementations.

goal criterion	best model	sensitivity (TPR)	specificity (TNR)	precision	accuracy	balanced accuracy	area under the curve
accuracy	LIBSVM BLAST-augm. $m = 7$, $C = 8$	41.13% (17.25%)	99.49% (1.08%)	93.81% (13.10%)	88.59% (4.69%)	70.31% (8.74%)	0.8188 (0.0983)
balanced accuracy	PSVM w. bal. BLAST-augm. $m = 8$, $C = 2$, $\varepsilon = 1.3$	73.82% (18.93%)	78.96% (11.83%)	46.49% (11.66%)	78.68% (7.51%)	76.39% (7.86%)	0.8202 (0.0859)

Table 4 | Overview of model selection results obtained by cross-validation on the entire dataset. Column 2 displays the best parameter settings in terms of accuracy (first row) and balanced accuracy (second row). Columns 3–6 display average performance measures over all 10 test folds for the models trained using these best parameter settings (standard deviations in parentheses).

The first setting, LIBSVM's C-SVM trained with BLAST-augmented data using the normalized coiled coil kernel with $m = 7$ and the penalty parameter $C = 8$, optimizes for (standard) accuracy. It leads to excellent specificity, but sacrifices some sensitivity (thereby resulting in lower balanced accuracy). The second setting, PSVM trained with BLAST-augmented data using the normalized coiled coil kernel with m = 8 and regularization parameters $C = 2$ and $\varepsilon = 1.3$ (using balancing), achieves a better balance between sensitivity and specificity, thereby optimizing balanced accuracy, however, resulting in lower standard accuracy. It must be mentioned that both models have a consistently high ranking performance (as can be seen from the area-under-the-curve (AUC) values). Finally, two SVM models were trained according to the above settings with the entire BLAST-augmented dataset.

The classification results obtained are exceptionally good: The best model, hereafter called PrOCoil model, classified test (i.e., unknown) sequences with 88.6% accuracy, even though they had only a maximum identity of 60% to any (known) coiled coil with which the SVM was trained. This is especially remarkable, since this approach was not tested with only a few individual samples, but with all structurally resolved coiled coil sequences. This model is also used by the PrOCoil web tool (the R-implementation of PrOCoil also provides the PSVM-based model for advanced users).

In accordance with the stringent state-of-the-art testing methods employed, it can thus be concluded that new coiled coil samples can be classified with outstanding accuracy. The excellence of the classification results ensures that the rules subsequently extracted from this model are indeed based on significant patterns. The calculations show that classification was enhanced by training the SVM with the augmented dataset. By using only structurally resolved PDB sequences in the test datasets, it was made sure, however, that this improvement was not due to the optimization of an "artificial" dataset.

Pairwise patterns | 4.3.4

Based on the rules constructed by the coiled coil kernel approach, amino acid patterns were extracted from the augmented dataset. These patterns comprise pairs of amino acids at certain heptad positions that are characteristic of each type of oligomer (see **Figures 18** and **A8**). This information was then used to implement a prediction and sequence profiling tool, PrOCoil, that characterizes the overall oligomeric tendency of a coiled coil sequence by displaying each amino acid's contribution to the rules in which it participates in a specific sequence.

Although the statistical approach identifies similar individual amino acids as important, the new method presented in this work is able to provide a much more detailed picture of the influence of each amino acid on the overall structure by taking its neighborhood into account.

An amino acid at a certain position participates in various patterns; consequently, the extracted patterns are correlated. The factorization performed by PrOCoil is therefore essential to decorrelate the patterns and to reduce a coiled-coil-spanning network to its building blocks. The 25 most influential amino acid pairings according to the PrOCoil model are shown in **Figure 18**.

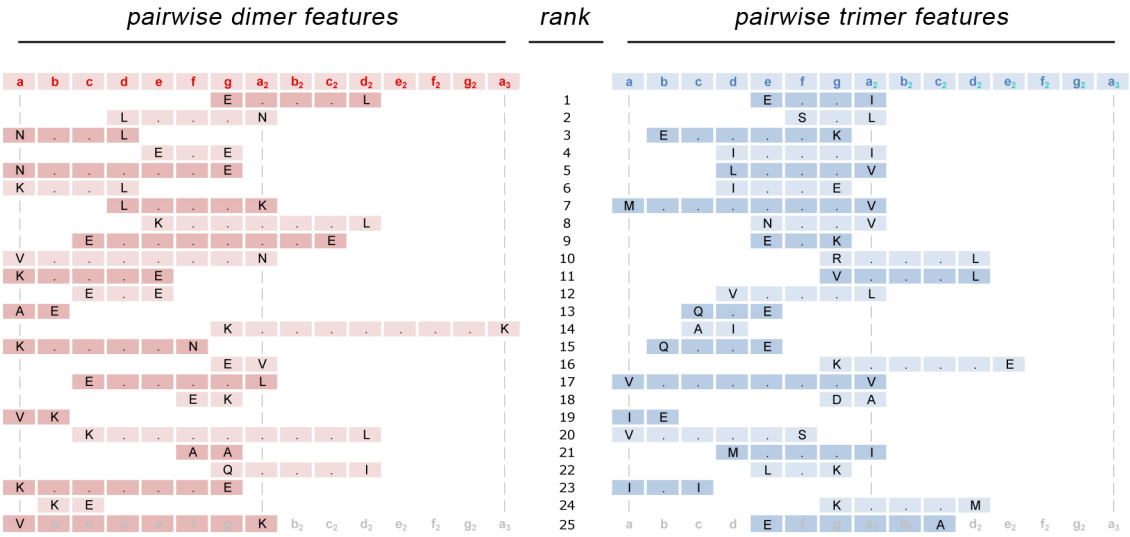


Figure 18 | List of the 25 strongest pairwise patterns. Dimer patterns are highlighted in pink, and trimer patterns are highlighted in blue. For instance, the top dimer pattern E...L, spanning columns g to d_2, describes a pattern with Glu at a *g* position, Leu at the next *d* position, and three arbitrary amino acids in between.

On closer inspection, this supports general hypotheses of oligomerization that rely on defining the hydrophobic core positions *a* and *d* [Harbury *et al.*, 1993] and can be extended to the mainly GCN4-leucin-zipper-based knowledge concerning the core positions to the whole heptad. I…I*d*, for example, is a well known trimer pattern of Ile at core positions *d* and *a* (with three arbitrary amino acids in between) that also ranks high in the shown list (as number 4). However, patterns that combine a core position with a non-core position seem to be at least as important to trimerization. For example, patterns with Leu, Ile, or Val at a core position and a tiny Ser at *f* position rank as numbers 2, 20 (**Figure 18**), and 75 (**Figure A8**), respectively. Ile at a core position combined with a charged Glu at *e* position ranks as number 1 and with Glu at *g* position as number 6.

The influence of the *b*, *c*, and *f* positions on oligomerization has long been underestimated because research has focused mainly on the positions in the hydrophobic core. These results, however, indicate that all positions inside a heptad contribute to the oligomeric tendency of a

coiled coil sequence. In fact, 9 out of the 25 most influential trimer patterns and 10 out of the 25 most influential dimer patterns are pairings with non-core positions (**Figure 18**).

In dimers, the highest-ranking patterns are combinations of amino acids with the β-branched Leu in core positions *a* and *d*. Interestingly, high-ranking β-branched combinations with Ile in a core position also occur in dimers (e.g., pattern 22). The fact that a single amino acid approach is insufficient because individual positions must always be viewed in context becomes particularly obvious when comparing the respective patterns ranked as number 2 for dimers and 5 for trimers. Both patterns have a Leu at *d* position, but when combined with Asn at *a* position (L...N*d*), it counts in favor of dimers, whereas with Val in *a* position (L…V*d*) it is characteristic of trimers. The 100 most important pairwise patterns for each oligomer according to the PrOCoil model can be found in the appendix of this work as **Figure A8**. It has been observed [Conway *et al.*, 1990, 1991] that a proportion of charged residues is absent in the hydrophobic core of three-stranded coiled coils. Investigation of the 100 most important patterns confirms these findings: Dimeric sequences have many patterns with charged amino acids inside the hydrophobic core. Positively charged amino acids are absent from the hydrophobic core of trimers.

While basic residues (His, Lys, Arg) in *a* position can be found in 16 patterns and acidic residues (Asn and Gln) in both *a* and *d* positions can be found in 12 of the 100 most important pairwise patterns in dimers (**Figure A8**), they are apparently not favored in trimers: only 3 of the top 100 trimer patterns feature acidic residues in the core, and none contain basic amino acids in these positions. According to Lupas *et al.* [2005], these observations can be attributed to the increased size and decreased solubility of the hydrophobic core in three-stranded structures, as well as to the acute packing orientation of core residues.

Sequence profiling | 4.3.5

As previously mentioned, PrOCoil can be used to visualize each amino acid's contribution to the oligomeric tendency of a sequence in a profiling plot. The following figures depict such PrOCoil sequence profiling plots using the example of a typical dimer, c-Jun (**Figure 19**), and a typical trimer, hemagglutinin, (**Figure 20**).

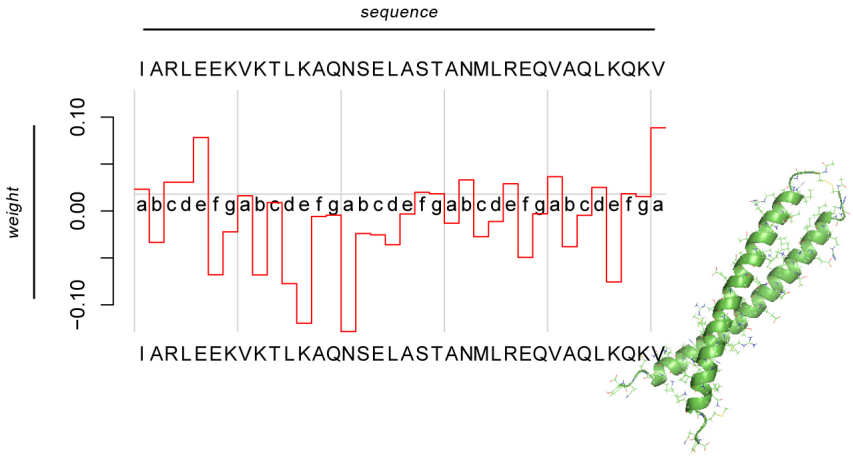


Figure 19 | Sequence profiling and classification of a typical dimer. The plot shows the c-Jun/c-Jun AP1 dimer, based on pairwise patterns of the PrOCoil model and visualizes the contribution of each amino acid position to the overall oligomeric tendency. The area above the base line equates to the positive/trimeric contributions, the area below corresponds to the negative/dimeric contributions. The 3D-figure was created using PyMOL (http://pymol.sourceforge.net/).

The sequences show a clear overall dimeric and trimeric tendency respectively. These tendencies are indicated by the areas above and below the base line, which equate respectively to the positive/trimeric contributions, and the negative/dimeric contributions of the involved patterns. The assigned weights are based on pattern-networks as the example in **Figure 21A** on the following page shows.

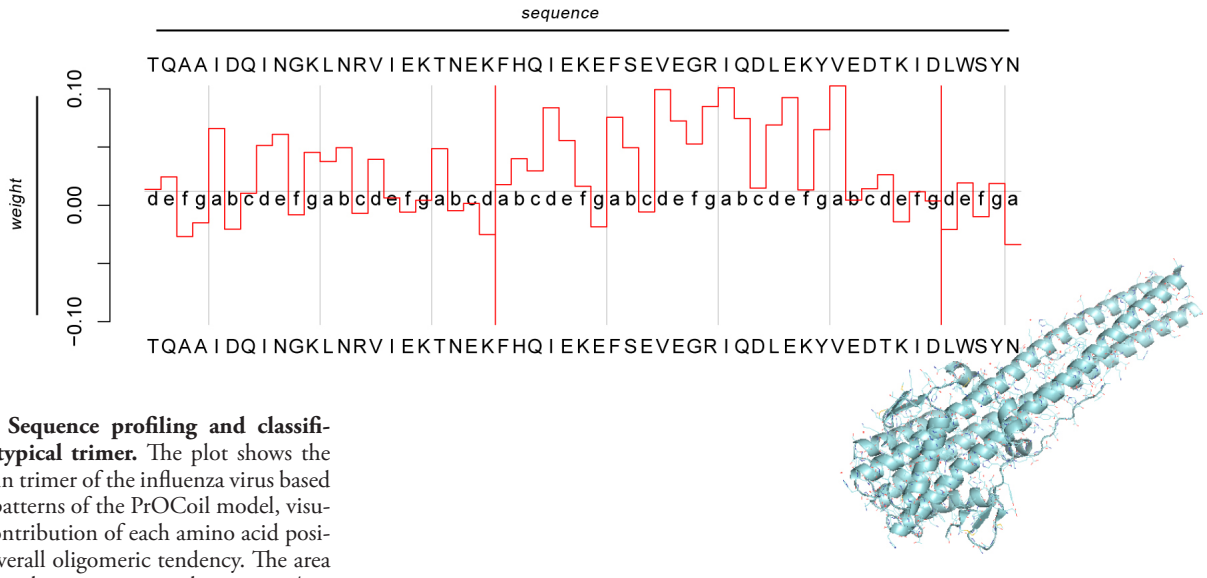


Figure 20 | Sequence profiling and classification of a typical trimer. The plot shows the hemagglutinin trimer of the influenza virus based on pairwise patterns of the PrOCoil model, visualizing the contribution of each amino acid position to the overall oligomeric tendency. The area above the base line equates to the positive/trimeric contributions, the area below corresponds to the negative/dimeric contributions. The 3D-figure was created using PyMOL (http://pymol.sourceforge.net/).

Figure 21B depicts the same kind of plot for wildtype GCN4 – a dimeric coiled coil protein renowned for its ability to adopt easily a different oligomerization state with very few mutations of the amino acid sequence [Portwich *et al.*, 2007]. The plot shows how this is possible – unlike c-Jun and hemagglutinin, GCN4 does not display a clear tendency towards one oligomeric state. It is impossible to assess at first glance whether the area above or below the base line is larger. As this protein combines both dimeric and trimeric characteristics (**Figure 21A**), it takes only a few selected mutations to tip the scales in one direction or the other.

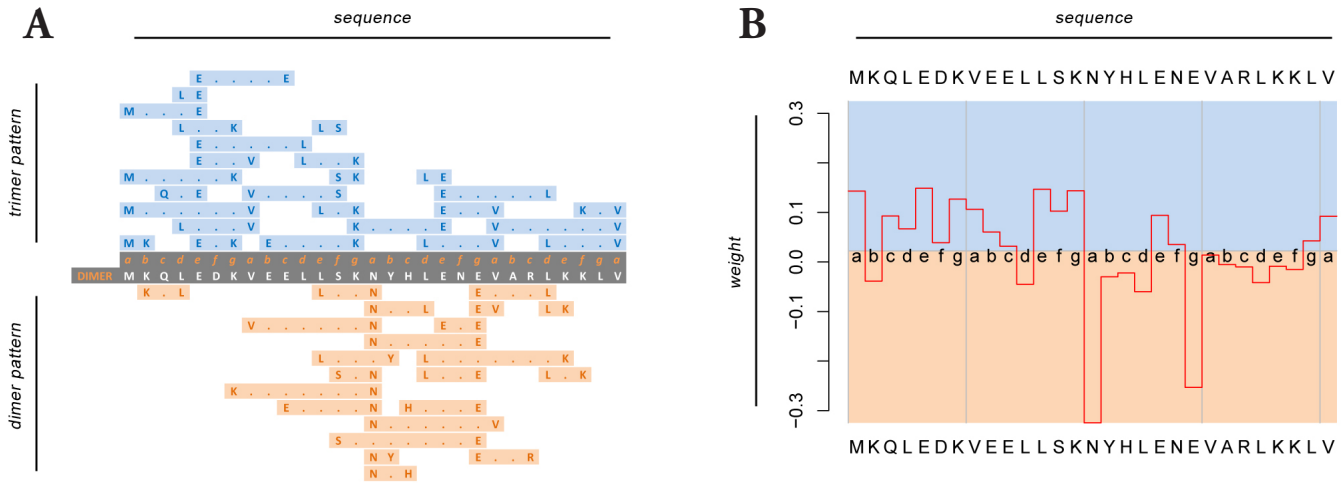


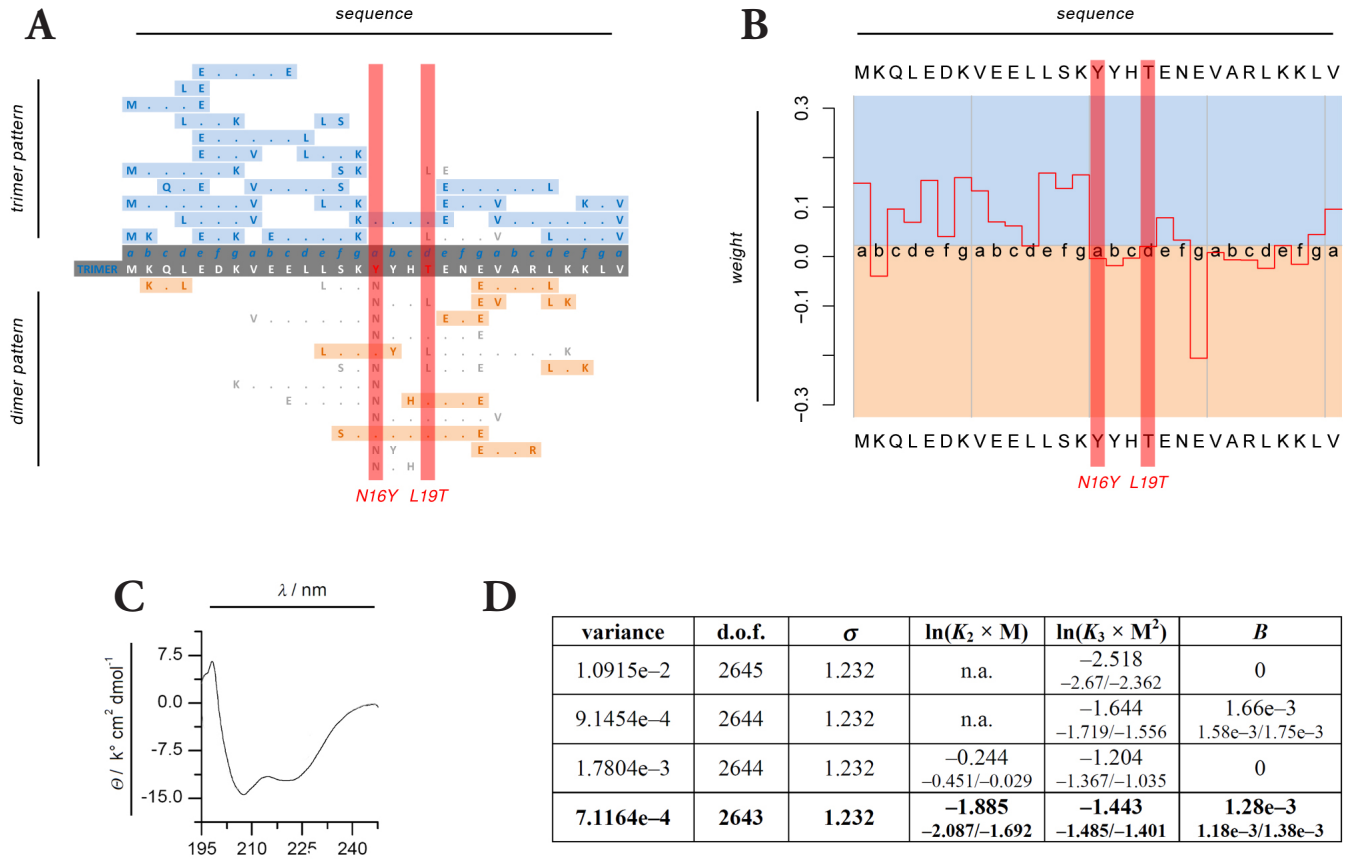
Figure 21 | Sequence profiling and classification of GCN4$_{wt}$. The plots show the dimeric transcriptional activator protein GCN4$_{wt}$. **(A)** The top pairwise patterns (listed in the appendix of this work, see Figure A8) found in GCN4$_{wt}$ are depicted on the left side to visualize which patterns are part of the heptad-spanning network. **(B)** The sequence-profiling plot on the right side visualizes the contribution of each amino acid position to the overall oligomeric tendency, based on the pairwise patterns from the PrOCoil model. The area above the base line (blue) equates to the positive/trimeric contributions, the area below (orange) corresponds to the negative/dimeric contributions.

GCN4 was thus the ideal candidate to demonstrate that the PrOCoil model can not only be used to provide excellent classification of wildtype sequences, but can even be employed for mutation analysis, given that a sufficiently large set of (similar) samples is provided with which it can be trained.

Mutation analysis of GCN4 mutants using PrOCoil | 4.3.6

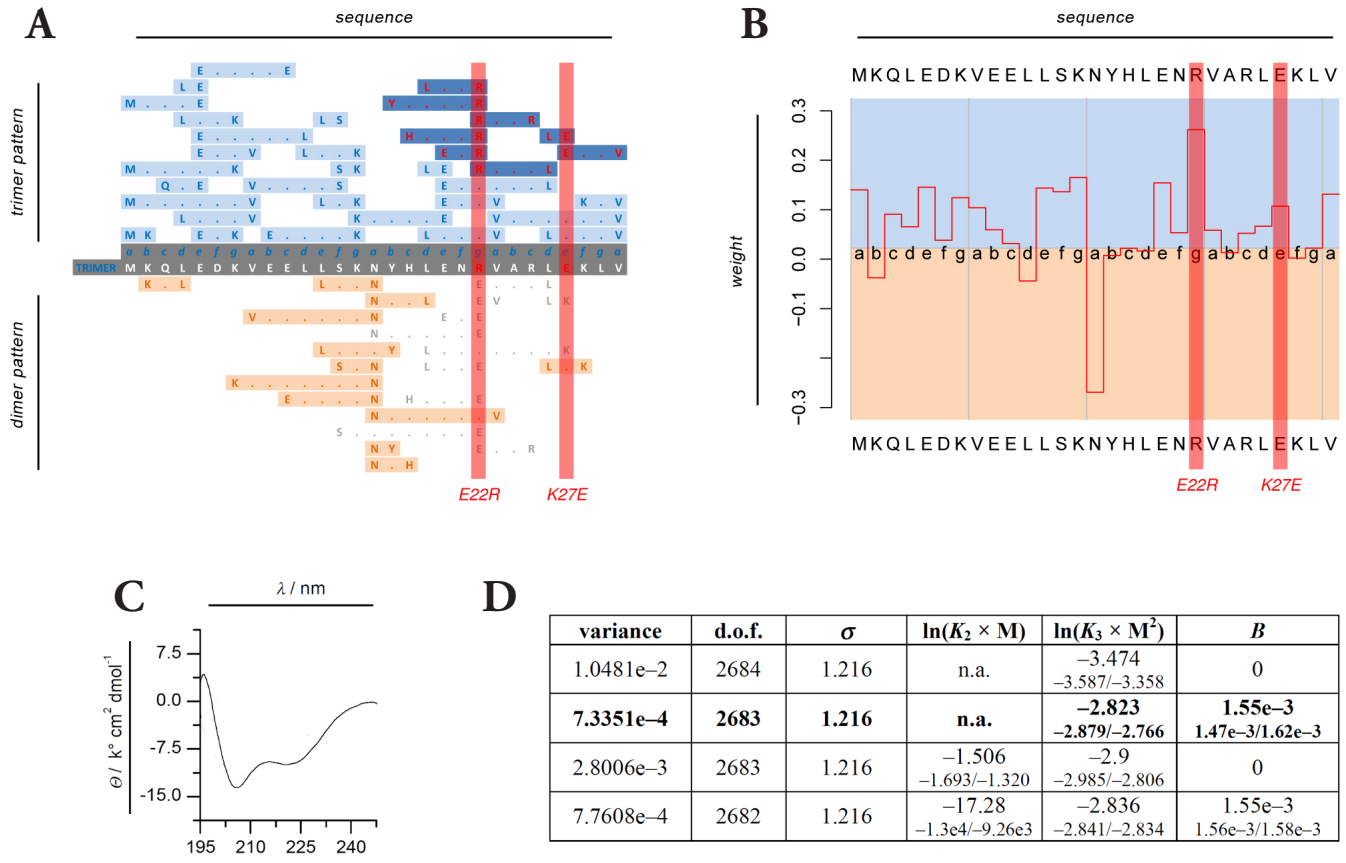
Two aforementioned mutant GCN4 sequences were chosen from the double-substitution analysis: $GCN4_{N16Y,L19T}$ and $GCN4_{V23K,K27E}$. Their oligomeric states and that of a sample which was mutated using the trimerizer pattern Arg(*g*)-h(*a*)-x-x-h(*d*)-Glu(*e*) [Kammerer *et al.*, 2005] were assessed by analytical ultra centrifugation. These mutants were not part of the datasets used for pattern extraction. Predicting their oligomerization states is challenging because the amino acid sequences of the dimeric wildtype GCN4 and its dimeric and trimeric mutants differ only at very few positions. PrOCoil classified all samples correctly, as visualized by the corresponding profiling plots in **Figures 22–24**. Additionally, the changes caused by the mutations in the pairwise patterns are shown. It immediately becomes apparent that the Asn in position 16 contributes the most weight to the dimeric tendency of $GCN4_{wt}$ (see **Figure 21**). This Asn participates in ten of the top 100 dimer patterns (see **Figure A8**), notably also in two of the three strongest of them all: L...N*d* and N..L*a*. From this follows that Asn16 represents the ideal target for mutation analysis to switch oligomerization (**Figure 22**).

Figure 22 | Sequence profiling and classification of $GCN4_{N16Y,L19T}$. The plots show the trimeric $GCN4_{N16Y,L19T}$ mutant. **(A)** The top pairwise patterns (listed in the appendix of this work, see Figure A8) found in $GCN4_{N16Y,L19T}$ are depicted on the left side to visualize which patterns are part of the heptad-spanning network. The color coding also depicts which patterns are added (dark color) or lost (grey) due to mutation. **(B)** The sequence-profiling plot on the right side visualizes the contribution of each amino acid position to the overall oligomeric tendency, based on the pairwise patterns from the PrOCoil model. The area above the base line (blue) equates to the positive/trimeric contributions, the area below (orange) corresponds to the negative/dimeric contributions. Red bars mark the positions that were mutated. **(C)** Corresponding circular-dichroism analysis of the mutant. The mean residue ellipticity, Θ, is plotted versus the wavelength, λ. The variant shows a clear α-helical tendency with minima at 208 and 222. **(D)** Fitting results returned by different homo-association models of the trimeric $GCN4_{N16Y,L19T}$ mutant as reported by NONLIN. Values are given in fringe units rather than molar quantities. PSV: 0.703 $mL \cdot g^{-1}$, ρ: 1.007 $g \cdot mL^{-1}$, Mw: 3748 Da, $\sigma_{50krpm,\ theoretical}$: 1.232. Best-fit values are shown in bold (d.o.f.: degrees of freedom, n.a.: not applicable).



variance	d.o.f.	σ	ln(K_2 × M)	ln(K_3 × M^2)	*B*
1.0915e–2	2645	1.232	n.a.	–2.518 –2.67/–2.362	0
9.1454e–4	2644	1.232	n.a.	–1.644 –1.719/–1.556	1.66e–3 1.58e–3/1.75e–3
1.7804e–3	2644	1.232	–0.244 –0.451/–0.029	–1.204 –1.367/–1.035	0
7.1164e–4	**2643**	**1.232**	**–1.885** **–2.087/–1.692**	**–1.443** **–1.485/–1.401**	**1.28e–3** **1.18e–3/1.38e–3**

Simple deletion of core-stabilizing dimer patterns by replacing Asn16 with Tyr and Leu19 with Thr while not adding important trimeric patterns results in a correctly predicted change in oligomerization (**Figure 22**). That is, the overall effect of destroying strong dimer patterns is sufficient to tip the oligomeric tendency in the direction of trimerization.



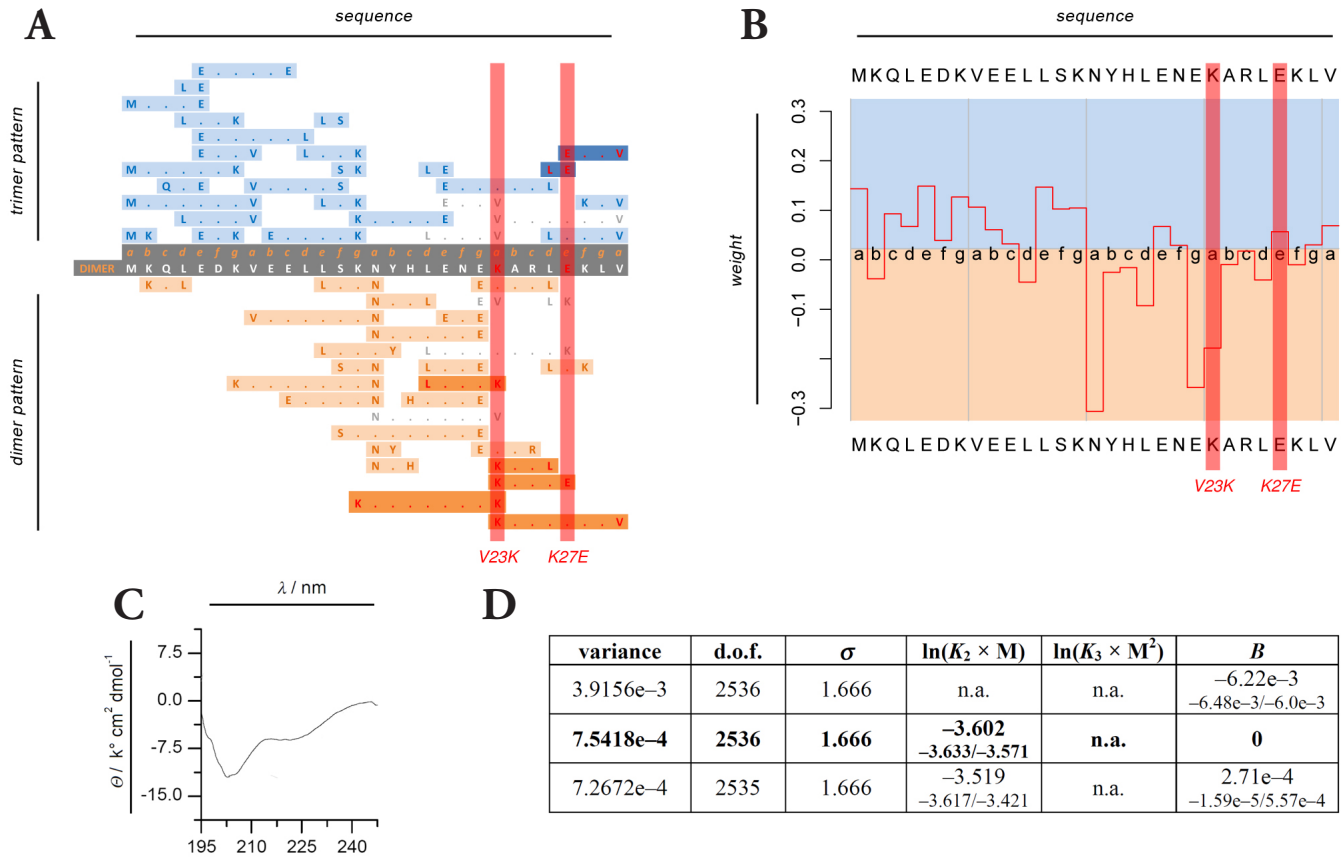
variance	d.o.f.	σ	ln(K_2 × M)	ln(K_3 × M^2)	B
1.0481e–2	2684	1.216	n.a.	–3.474 –3.587/–3.358	0
7.3351e–4	**2683**	**1.216**	**n.a.**	**–2.823** **–2.879/–2.766**	**1.55e–3** **1.47e–3/1.62e–3**
2.8006e–3	2683	1.216	–1.506 –1.693/–1.320	–2.9 –2.985/–2.806	0
7.7608e–4	2682	1.216	–17.28 –1.3e4/–9.26e3	–2.836 –2.841/–2.834	1.55e–3 1.56e–3/1.58e–3

Figure 23 | Sequence profiling and classification of GCN4$_{E22R,K27E}$. The plots show the trimeric GCN4$_{E22R,K27E}$ mutant. **(A)** The top pairwise patterns (listed in the appendix of this work, see Figure A8) found in GCN4$_{E22R,K27E}$ are depicted on the left side to visualize which patterns are part of the heptad-spanning network. The color coding also depicts which patterns are added (dark color) or lost (grey) due to mutation. **(B)** The sequence profiling plot on the right side visualizes the contribution of each amino acid position to the overall oligomeric tendency, based on the pairwise patterns from the PrOCoil model. The area above the base line (blue) equates to the positive/trimeric contributions, the area below (orange) corresponds to the negative/dimeric contributions. Red bars mark the positions that were mutated. **(C)** Corresponding circular-dichroism analysis of the mutant. The mean residue ellipticity, Θ, is plotted versus the wavelength, λ. The variant shows a clear α-helical tendency with minima at 208 and 222. **(D)** Fitting results returned by different homo-association models of the trimeric GCN4$_{E22R,K27E}$ mutant as reported by NONLIN. Values are given in fringe units rather than molar quantities. PSV: 0.704 mL·g^{-1}, ρ: 1.007 g·mL^{-1}, Mw: 3711 Da, $\sigma_{50krpm,\ theoretical}$: 1.216. Best-fit values are shown in bold (d.o.f.: degrees of freedom, n.a.: not applicable).

The next example (**Figure 23**) shows that, alternatively, trimerization can also be triggered without replacing Asn16 if strong trimer patterns are added instead. For this purpose, the trimerizer pattern [Kammerer *et al.*, 2005] was inserted.

As verified experimentally, and PrOCoil predicted correctly, oligomerization switches if mutations are inserted at positions that create strong trimer patterns. Arbitrary mutations in this sequence region do not change oligomerization. The substitutions must be selected carefully to create additional trimer patterns. To prove this point, a sequence was chosen in which Val23 was replaced by Lys, and Lys27 by Glu (**Figure 24** on the following page). This resulted in the loss of three important Val23-related trimer patterns, while only two Lys-related ones were added. At the same time, a strong Lys- and Glu-related and four strong Lys-related dimeric patterns were created. This substitution with

physico-chemically similar amino acids added new dimer patterns and, as predicted, the mutations do not affect oligomerization.



variance	d.o.f.	σ	$\ln(K_2 \times M)$	$\ln(K_3 \times M^2)$	B
3.9156e–3	2536	1.666	n.a.	n.a.	–6.22e–3 –6.48e–3/–6.0e–3
7.5418e–4	**2536**	**1.666**	**–3.602** **–3.633/–3.571**	**n.a.**	**0**
7.2672e–4	2535	1.666	–3.519 –3.617/–3.421	n.a.	2.71e–4 –1.59e–5/5.57e–4

Figure 24 | Sequence profiling and classification of GCN4$_{V23K,K27E}$. The plots show the trimeric GCN4$_{V23K,K27E}$ mutant. **(A)** The top pairwise patterns (listed in the appendix of this work, see Figure A8) found in GCN4$_{V23K,K27E}$ are depicted on the left side to visualize which patterns are part of the heptad-spanning network. The color coding also depicts which patterns are added (dark color) or lost (grey) due to mutation. **(B)** The sequence profiling plot on the right side visualizes the contribution of each amino acid position to the overall oligomeric tendency, based on the pairwise patterns from the PrOCoil model. The area above the base line (blue) equates to the positive/trimeric contributions, the area below (orange) corresponds to the negative/dimeric contributions. Red bars mark the positions that were mutated. **(C)** Corresponding circular-dichroism analysis of the mutant. The mean residue ellipticity, Θ, is plotted versus the wavelength, λ. The variant shows a clear α-helical tendency with minima at 208 and 222. **(D)** Fitting results returned by different homoassociation models of the trimeric GCN4$_{V23K,K27E}$ mutant as reported by NONLIN. Values are given in fringe units rather than molar quantities. PSV: 0.720 mL·g^{-1}, ρ: 1.007 g·mL^{-1}, Mw: 3741 Da, $\sigma_{50krpm,\ theoretical}$: 1.666. Best-fit values are shown in bold (d.o.f.: degrees of freedom, n.a.: not applicable).

The examples show: The influence of mutations on oligomerization depends on the sequence context, i.e., on the overall effect of the change in the interdependent patterns caused by the mutations. Added trimer patterns and/or loss of strong dimer patterns results in an overall trimeric structure, whereas adding strong dimer patterns maintains dimerization of the protein.

Conclusions

Analysis of three common detection agents to reveal their ability to cross-react with cellulose-bound peptides.

A methodical contribution of this work was to demonstrate that several amino acids interact with TAMRA-, FITC-, or biotin-labeling agents and streptavidin-POD. FITC cannot be recommended for read-out when probing peptide arrays on cellulose membranes for binding, and taking these results into consideration is also advisable for *in vivo* approaches. Besides spot autofluorescence of several amino acids and the cellulose membrane, label-specific cross-reactivity with positively charged amino acids was observed for FITC and streptavidin-POD. TAMRA, in contrast, seems to be a more suitable screening/read-out system for probing peptide arrays on a cellulose membrane. However, the influence of aromatic amino acids, especially tryptophan, must be taken into account. Critical examination of peptide sequences is essential, and a comparative approach using both TAMRA- and biotin-labeled analytes is recommended. Such an approach compensates for effects on label-specific cross-reactive amino acids.

One has to bear in mind that a method is always limited by the effectiveness and validity of the read-out system. To avoid unwanted side effects, the right choice of buffer solutions is advised. However, in the case of SPOT synthesis, even the optimal buffer composition [Beutling *et al.*, 2008; Bräuning *et al.*, 2002] fails to prevent cross-reaction. This work identified several amino acids that interact with different detection systems. To prevent or identify false positives, factoring in these results is highly recommended when analyzing measurements.

Furthermore, as good experimental practice [Frank *et al.*, 1996], testing the detection method of choice for its ability to cross-react before running the actual experiment is strongly advised. Taking these new results into consideration will, in future, strengthen the reliability of the analysis of SPOT-synthesis-generated data.

Analysis of coiled coils using peptide libraries.

In conclusion, the results presented here demonstrate that synthetic peptide arrays, despite their limitations, can be used for studying coiled coil associations – even for distinguishing coiled-coil from other interactions, based on their unique substitution patterns. However, they do not provide information about the stoichiometry and topology of the coiled coils. Due to the fact that peptide array binding studies are limited to the analysis of peptide associations, the only way to uncover coiled coil oligomerization in detail was by combining biochemical technologies (i.e., SPOT- and peptide synthesis) with quantitative biophysical measurements of selected candidates. This experimental approach, however, is clearly insufficient to uncover the complex hidden rules of oligomerization, Revealing them requires a more comprehensive approach.

Development and testing of a new sequence-based theory to explain coiled coil oligomerization.

Although science has been seeking to explain coiled coil behavior for decades, the majority of current approaches considered state of the art are based on analyzing only the influence of single amino acids in isolation and deliver limited or incomplete explanations. Disregarding the sequence context is – as shown in this work – a result of oversimplifying a complex phenomenon. Inspired by approaches that sought to define and predict coiled coils based on statistics, the first attempt at finding the rules for dimeric and trimeric oligomerization was to examine the position-specific single amino acid frequencies in each oligomer. In contrast to hitherto published coiled coil statistics [e.g. Lupas *et al.*, 1991, Gruber *et al.*, 2006], the single amino acid statistics are based on clustered data. This compensates for the (artificially) high prevalence of certain sequences in the PDB stemming from concentrated scientific interest in certain types of proteins. However, the results show that a simple statistical analysis cannot provide an explanation or rules for a sequence's preference for a certain oligomeric state.

As demonstrated in this work, stepping up to a higher level of complex-

ity, examining the relations between amino acids, is the key to predicting and understanding oligomerization. For the first time, a complete network of sequence parameters that influence oligomerization has been established, and the underlying rules of coiled-coil formation have been provided. Support Vector Machines using the coiled coil kernel as the method of choice make it possible to classify with outstanding accuracy dimers and trimers from their amino acid sequences, and to obtain the valuable rules (weighted patterns) learned by the machine to determine oligomeric preference. Each individual pattern is, however, only a part of the whole formula that explains coiled coil oligomerization. The patterns must be viewed in context, as parts of a network of interactions, to understand and predict oligomerization. Using the example of GCN4, it was demonstrated that this information can be merged to design the sequence analyzing tool PrOCoil and draw an overall picture that explains the behavior of a sequence that, until now, seemed unclassifiable. Moreover, PrOCoil can even be used to indicate which sequence positions contribute most to the dimeric or trimeric tendency, and it provides exceptionally reliable sequence-based prediction of coiled coil stoichiometry. To ensure that the conclusions about oligomerization are valid for coiled coils in general and not limited to a few representative cases, as has often been the case in this field of research, this approach is based on, and verified using, all structurally resolved coiled coils. The software (PrOCoil) developed in cooperation with the Institute of Bioinformatics (Linz, Austria) in the course of this work will provide the scientific community with a powerful tool that both classifies and visualizes the oligomeric tendency of a coiled coil sequence at single amino acid resolution, which is particularly useful for mutation analysis and *de novo* protein design. PrOCoil is available as a web-based tool and already used by selected scientists. A web version and an R package of the prediction and profiling software (PrOCoil) are available to the scientific community (see http://www.bioinf.jku.at/software/procoil/).

In summary, all initial postulates were verified and led to a refined understanding of coiled coil behavior doing justice to their complexity. The data collected and the tool designed in this work offer a new basis for coiled coil prediction and design. These findings at last elucidate the link between coiled coil sequence and structure.

Closing remarks

Next generation sequencing techniques will soon provide us with a vast quantity of new sequence data. This is of great interest in the context of this work, since the BLAST approach described here can be used to tap this source of data for a wealth of new training sequences. Subsequent retraining of SVMs with the coiled coil kernel should further refine the pattern set, lead to new insights, and further improve the already excellent prediction performance of PrOCoil.

It was postulated in chapter 1 that coiled coil oligomerization is poorly understood. This assertion must, in conclusion, be revisited and the question asked again whether we are now in a position to draw a clearer picture of coiled coil oligomerization.

Answering this question necessitates finding a philosophical approach that is complementary to the natural sciences. As scientists, we seek to model natural phenomena, approximating reality with theory. Sometimes we succeed in making nature comprehensible, predictable, and even representable mathematically. However, even when the resulting theoretical impression of nature seems plausible, we must never lose sight of the fact that that we are dealing with simplified abstractions. We seek to interpret aspects of complex phenomena by means of experiments and with the help of models. Our situation is similar to that of the prisoners in Plato's cave allegory: Chained in a cave, they can watch

only the shadows cast on the wall by things that pass the fire behind them [see Platon, 370BC]. The shadows are as close as they get to perceiving reality.

Hence an answer to the initial question must take into account the technical limitations of the state of the art: researchers can see only what technology (in this case fire) can make visible, and subjective reality (an observed shadow) is but a derivative of objective reality, perceived by technological means, and subsequently interpreted. This intrinsic uncertainty can only strengthen the imperative that the technical means by which we explore and seek to understand nature be as reliable as possible.

The answer to the underlying question of what we know must therefore be that while, on the one hand, this work advances significantly the state of the art in its field, on the other hand, the future will doubtless bring further refinements and insights.

Bibliography

A

Altschul, S.F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Andresen, H. & Bier, F.F. Peptide microarrays for serum antibody diagnostics. *Methods Mol. Biol.* **509**, 123-134 (2009).

B

Behncken, S.N., *et al.* Growth hormone(GH)-independent dimerization of GH receptor by a leucine zipper results in constitutive activation. *J. Biol. Chem.* **275**, 17000-7 (2000).

Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**(1), 289–300 (1995).

Berger, B. & Singh, M. An iterative method for improved protein structural motif recognition. *J. Comput. Biol.* **4**, 261-273. (1997)

Berger, B. *et al.*, Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl. Acad. Sci. USA.* **92**, 8259–8263 (1995).

Bernstein, F.C. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**(3), 535–542 (1977).

Beutling, U., Städing, K., Stradal, T. & Frank, R. Large-scale analysis of protein-protein interactions using cellulose-bound peptide arrays. *Adv. Biochem. Engin./Biotechnol.* **110**, 115-152 (2008)

Bianchi, E. *et al.* Covalent stabilization of coiled coils of the HIV gp41 N region yields extremely potent and broad inhibitors of viral infection. *Proc. Natl. Acad. Sci. USA.* **102**(36), 12903–8 (2005).

Bodenhofer, U., Schwarzbauer, K., Ionescu, M. & Hochreiter, S. *Modeling position specificity in sequence kernels by fuzzy equivalence relations.* in Proc. Joint 13th IFSA World Congress and 6th EUSFLAT Conference, eds Carvalho J.P., Dubois D., Kaymak U. & Sousa J.M.C. pp 1376–1381 (2009).

Bräuning, R. *et al.* Immobilized peptides to study protein-protein interactions - potential and pitfalls, in *Peptide Arrays on Membrane Supports: Synthesis and Applications* (Eds.: Koch J, Mahler M), Springer, Heidelberg, 2002, pp. 153-163.

Burges, C.J.M. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121–167 (1998).

Burkhard, P., Meier, M. & Lustig, A. Design of a minimal protein oligomerization domain, a structural approach. *Protein Science* **9**, 2294-2301 (2000).

Burkhard, P., Stetefeld, J. & Strelkov, S.V. Coiled coils: a highly versatile protein folding motif. *Trends Cell. Biol.* **11**(2), 82–88 (2001).

C

Chang, C.C. & Lin, C.H. *LIBSVM : a library for support vector machines.* (2001) Software available at (http://www.csie.ntu.edu.tw/~cjlin/libsvm)

Christianini, N. & Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge University Press, 2000).

Contegno, F., Cioce, M., Pelicci, P.G. & Minucci, S. Targeting protein inactivation through an oligomerization chain reaction. *Proc. Natl. Acad. Sci. USA.* **99**, 1865-9 (2002).

Conway, J. F. & Parry, D. A.D. Threestranded a-fibrous proteins: the heptad repeat and its implication for structure. *Int. J. Biol. Macromol.* **13**, 14-16 (1991).

Conway, J. F. & Parry, D.A.D. Structural features in the heptad substructure and longer range repeats of two- stranded a-fibrous proteins. *Int. J. Biol. Macromol.* **12**, 328-334 (1990).

Cortes, C. & Vapnik, V.N. Support vector networks. *Machine Learning* **20**, 273–297 (1986).

Crick, F.H.C. Is α-keratin a coiled coil? *Nature* **170**, 882–883 (1952).

D

Delorenzi, M. & Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **18**(4), 617-25 (2002).

Dürauer, A. *et al.* Evaluation of a sensitive detection method for peptide arrays prepared by SPOT synthesis. *Biochem. Biophys. Methods* **66**(1-3), 45-57 (2006).

F

Fisher, R.A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. Roy. Statist. Soc.* **85**(1), 87–94 (1922).

Fong, J.H., Keating, A.E. & Singh, M. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biology* **5**, R11 (2004).

Frank, R. & Overwin, H. SPOT-Synthesis: Epitope analysis with arrays of synthetic peptides prepared on cellulose membranes, in *Methods in Molecular Biology*, Vol. 66: Epitope Mapping Protocols (Ed.: Morris GE) The Humana Press Inc., Totowa, 1996, pp. 149-169.

Frank, R. SPOT-synthesis: An easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. *Tetrahedron* **48**, 9217-9232 (1992).

Frank, R. The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports – principles and applications. *J. Immunol. Methods* **267**(1), 13-26 (2002).

G

Ghosh, I., Hamilton, A.D. & Regan, L. Antiparallel leucine zipper-directed protein reassembly: application to the green fluorescent protein. *J. Am. Chem. Soc.* **122**, 5658-9 (2000).

Gingras, A.R. *et al.* The structure of the C-terminal actin-binding domain of talin. *EMBO J.* **27**(2), 458–469 (2008).

Goldenberg, D.M. Advancing role of radiolabeled antibodies in the therapy of cancer. *Cancer Immunol. Immunother.* **52**, 281-96 (2003).

Goodwin, D.A. & Meares, C.F. Advances in pretargeting biotechnology. *Biotechnol. Adv.* **19**, 435-50 (2001).

Gruber, M., Söding, J. & Lupas, A.N. Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.* **155**, 140-145 (2006).

H

Hadley, E.N., Testa, O.D., Woolfson, D.N. & Gellman, S.H. Preferred side-chain constellations at antiparallel coiled-coil interfaces. *Proc. Natl. Acad. Sci. USA.* **105**(2), 530– 535 (2008).

Harbury, P.B., Zhang, T., Kim, P.S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).

Harlow, E. & Lane, D. *Antibodies – A laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1988, pp. 319-358.

Hartl, M., Bader, A.G. & Bister, K. Molecular targets of the oncogenic transcription factor jun. *Curr. Cancer Drug Targets* **3**, 41-55 (2003).

Henikoff, S. & Henikoff, J.G. Amino Acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* **90**, 10915-10919 (1992).

Hochreiter, S. & Obermayer, K. Support vector machines for dyadic data, *Neural Comput.* **18**, 1472–1510 (2006).

Hoiczyk, E. *et al.* Structure and sequence analysis of Yersinia YadA and Moraxella UspAs reveal a novel class of adhesins. *EMBO J.* **19**, 5989-5999 (2000).

Hu, J.C., Newell, N.E., Tidor, B. & Sauer, R.T. Probing the roles of residues at the e and g positions of the GCN4 leucine zipper by combinatorial mutagenesis. *Protein Sci.* **2**(7), 1072–1084 (1993).

Hua, Q.X. *et al.* Diabetes-associated mutations in a beta-cell transcription factor destabilize an antiparallel „mini-zipper" in a dimerization interface. *Proc. Natl. Acad. Sci. USA.* **97**(5),1999-2004 (2000).

J

Johnson, M.L., Correria, J.J., Yphantis, D.A. & Halvorson, H.R. Analysis of data from the analytical ultracentrifuge by nonlinear least-squares techniques. *Biophys. J.* **36**, 575–88 (1981).

K

Kammerer, R.A. *et al.* A conserved trimerization motif controls the topology of short coiled coils. *Proc. Natl. Acad. Sci. USA.* **102**(39), 13891–13896 (2005).

Katz, B.Z. *et al.* Green fluorescent protein labeling of cytoskeletal structures--novel targeting approach based on leucine zippers. *Biotechniques* **25**, 298-304 (1998).

Knox, S.J. *et al.* Phase II trial of yttrium-90-DOTA-biotin pretargeted by NR-LU-10 antibody/streptavidin in patients with metastatic colon cancer. *Clin. Cancer. Res.* **6**, 406-14 (2000).

Kramer, A. *et al.* Spot-synthesis: observations and optimizations. *J. Pept. Res.* **54**(4), 319-327 (1999).

Kratky, O., Leopold, H. & Stabinger, H. The determination of the partial specific volume of proteins by the mechanical oscillator technique. *Methods Enzymol.* **27**, 98–110 (1993).

Kuksa, P., Huang, P.-H. & Pavlovic, V. *A fast, large-scale learning method for protein sequence classification.* in 8th Int. Workshop on Data Mining in Bioinformatics, pp 29–37 (2008).

L

Leslie, C. *et al.* Mismatch string kernels for discriminative protein classification. *Bioinformatics* **1**(1), 1-10 (2003).

Leslie, C. *et al. The spectrum kernel: a string kernel for SVM protein classification*, in Pacific Symposium on Biocomputing, Altman, R.B. *et al.*, Eds, pp. 566-575 (2002).

Lewis, W.G. & Smillie, L.B. The amino acid sequence of rabbit cardiac tropomyosin. *J. Biol. Chem.* **255**, 6854-9 (1980).

Li, X., Gu, W., Mohan, S. & Baylink, D.J. DNA microarrays: their use and misuse. *Microcirculation* **9**(1), 13-22 (2002).

Lupas, A., Van Dyke, M. & Stock, J. Predicting Coiled Coils from Protein Sequences. *Science* **252**, 1162-1164 (1991).

Lupas, A.N. & Gruber, M. The structure of a-helical coiled coils. *Adv. Protein Chem.* **70**, 37-78 (2005).

Lupas, A.N. The long coming of computational structural biology. *J. Struct. Biol.* **163**, 254–257 (2008).

M

Malashkevich, V.N. *et al.* Core structure of the envelope glycoprotein GP2 from Ebola virus at 1.9-A resolution. *Proc. Natl. Acad. Sci. USA.* **6**, 2662-7 (1999).

Mason, J.M., Müller, K.M. & Arndt, K.M. iPEP: peptides designed and selected for interfering with protein interaction and function. *Biochem. Soc. Trans.* **36**, 1442-1447 (2008).

McDonnell, A.V., Jiang, T., Keating, A.E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**(3), 356-8 (2006).

McFarlane, A.A. *et al.* The use of coiled-coil proteins in drug delivery systems. *Eur. J. Pharmacol.* **625**(1–3), 101–107 (2009).

McLachlan, A.D. & Karn, J. Periodic features in the amino acid sequence of nematode myosin rod. *J. Mol. Biol.* **164**, 605-626 (1983).

Merrifield, R.B., Solid phase peptide synthesis. I. The Synthesis of a Tetrapeptide. *J.Am.Chem.Soc.* **85**, 2149-2154 (1963).

Moll, J.R., Ruvinov, S.B., Pastan, I. & Vinson, C. Designed heterodimerizing leucine zippers with a ranger of pIs and stabilities up to 10-15 M. *Protein Sci.* **10**, 649-55 (2001).

Müller, K.R. *et al.* An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks* **12**(2), 181–201 (2001).

O

O'Shea, E.K., Lumb, K.J. & Kim, P.S. Peptide 'Velcro': design of a heterodimeric coiled coil. *Current Biology* **3**, 658–667 (1993).

Otte, L. *et al.* WW domain sequence activity relationships identified using ligand recognition propensities of 42 WW domains. *Protein Sci.* **12**, 491–500 (2003).

P

Pauling, L. & Corey, R.B. Compound helical configurations of polypeptide chains: Structure of proteins of the α-keratin type. *Nature* **171**, 59–61 (1953).

Petka, W.A. *et al.* Reversible hydrogels from self-assembling artificial proteins. *Science* **281**, 389-92 (1998).

Platon. *Politeia.* 370BC, pp. 514a-517a.

Portwich, M. *et al.* A network of coiled coil associations derived from synthetic GCN4 leucine zipper arrays. *Angew. Chem. Int. Ed.* **46**, 1654–1657 (2007).

Pritchard, C., Underhill, P. & Greenfield, A. Using DNA micro arrays. *Methods Mol. Biol.* **461**, 605-629 (2008).

R

Reimer, U., Reineke, U. & Schneider-Mergener, J. Peptide arrays: from macro to micro. *Curr. Opin. Biotechnol.* **13**(4), 315-320 (2002).

Reinecke, U., Volkmer-Engert, R. & Schneider-Mergener, Applications of peptide arrays prepared by the Spot technology. *J. Curr. Opin. Biotechnol.* **12**, 59-64, (2001).

Reineke, U., Schneider-Mergener, J. & Schutkowski, M. *Peptide array in proteomics and drug discovery, in BioMEMS and Biomedical Nanotechnology, Vol.II Micro and Nano-technologies for Genomics and Proteomics* (Ed.: M. Ozkan, M. J. Heller), Springer, New York, 2005, pp. 161-282.

Russell, R.J. *et al.* Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion. *Proc. Natl. Acad. Sci. USA* **105**(46), 17736–41 (2008).

S

Schölkopf, B., Tsuda, K. & Vert, J.P. Eds. *Kernel Methods in Computational Biology* (MIT Press, Cambridge, MA, 2004).

Sheriff, S., Chang, C.Y. & Ezekowitz, R.A. Human mannose-binding protein carbohydrate recognition domain trimerizes through a triple alpha-helical coiled-coil. *Nat. Struct. Biol.* **1**(11), 789–794 (1994).

Stoevesandt, O., Taussig, M.J. & He, M. Protein microarrays: high-throughput tools for proteomics. *Expert Rev. Proteomics* **6**(2), 145-157 (2009).

Strauss, H.M. & Keller, S. Pharmacological interference with protein-protein interactions mediated by coiled-coil motifs. *Handb. Exp. Pharmacol.* **186**, 461-82 (2008).

Sygula, A. *et al.* A double concave hydrocarbon buckycatcher. *J. Am. Chem. Soc.* **129**(13), 3842–3843 (2007).

T

Tang, A., Wang, C., Stewart, R.J. & Kopecek, J. The coiled coils in the design of protein-based constructs: hybrid hydrogels and epitope displays. *J. Control Release* **72**, 57-70 (2001).

Tao, S.C., Chen, C.S. & Zhu, H. Applications of protein microarray technology. *Comb. Chem. High Throughput Screen.* **10**(8), 706-718 (2007).

Toepert, F. *et al.* Combining SPOT synthesis and native peptide ligation to create large arrays of WW domains. *Angew. Chem. Int. Ed.* **42**(10), 1136-1140 (2003).

Turner, M.W. Mannose-binding lectin (MBL) in health and disease. *Immunobiology* **199**(2), 327-39 (1998).

U

Uttamchandani, M., Li, J., Sun, H. & Yao, S.Q. Activity-based protein profiling: new developments and directions in functional proteomics. *ChemBioChem* **9**(5), 667-675 (2008).

V

Vapnik, V.N. *Statistical Learning Theory. Adaptive and Learning Systems.* (Wiley New York, 1998).

Volkmer, R. Synthesis and application of peptide arrays: quo vadis SPOT technology. *ChemBioChem* **15**(9), 1431-1442 (2009).

W

Walshaw, J. & Woolfson, D.N. SOCKET: A program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.* **307**, 1427–1450 (2001).

Wang, C, Stewart, R.J. & Kopecek, J. Hybrid hydrogels assembled from synthetic polymers and coiled-coil protein domains. *Nature* **397**, 417-20 (1999).

Wang, C., Kopecek, J. & Stewart, R.J. Hybrid hydrogels cross-linked by genetically engineered coiled-coil block proteins. *Biomacromolecules* **2**, 912-20 (2001).

Weiser, A.A. *et al.* SPOT synthesis: reliability of array-based measurement of peptide binding affinity. *Anal. Biochem.* **342**(2), 300-311 (2005).

Wenschuh, H. *et al.* Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Biopolymers* **55**(3), 188-200 (2000).

Wilson, M.B. & Nakane, P.K. *Immunofluorescence related staining techniques.* (Eds.: W. Knapp, K. Holubar, G. Wiek), Elsevier, Amsterdam, 1978, pp. 215-224.

Winkler, D.F.H. & McGeer, P.L. Protein labeling and biotinylation of peptides during spot synthesis using biotin p-nitrophenyl ester (biotin-ONp). *Proteomics* **8**, 961–967 (2008).

Wolf, E., Kim, P.S. & Berger, B. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* **6**, 1179–1189 (1997).

Woolfson, D.N. & Alber, T. Predicting oligomerization states of coiled coils. *Protein Sci.* **4**, 1596–1607 (1995).

Z

Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**(5537), 2101-2105 (2001).

Appendix

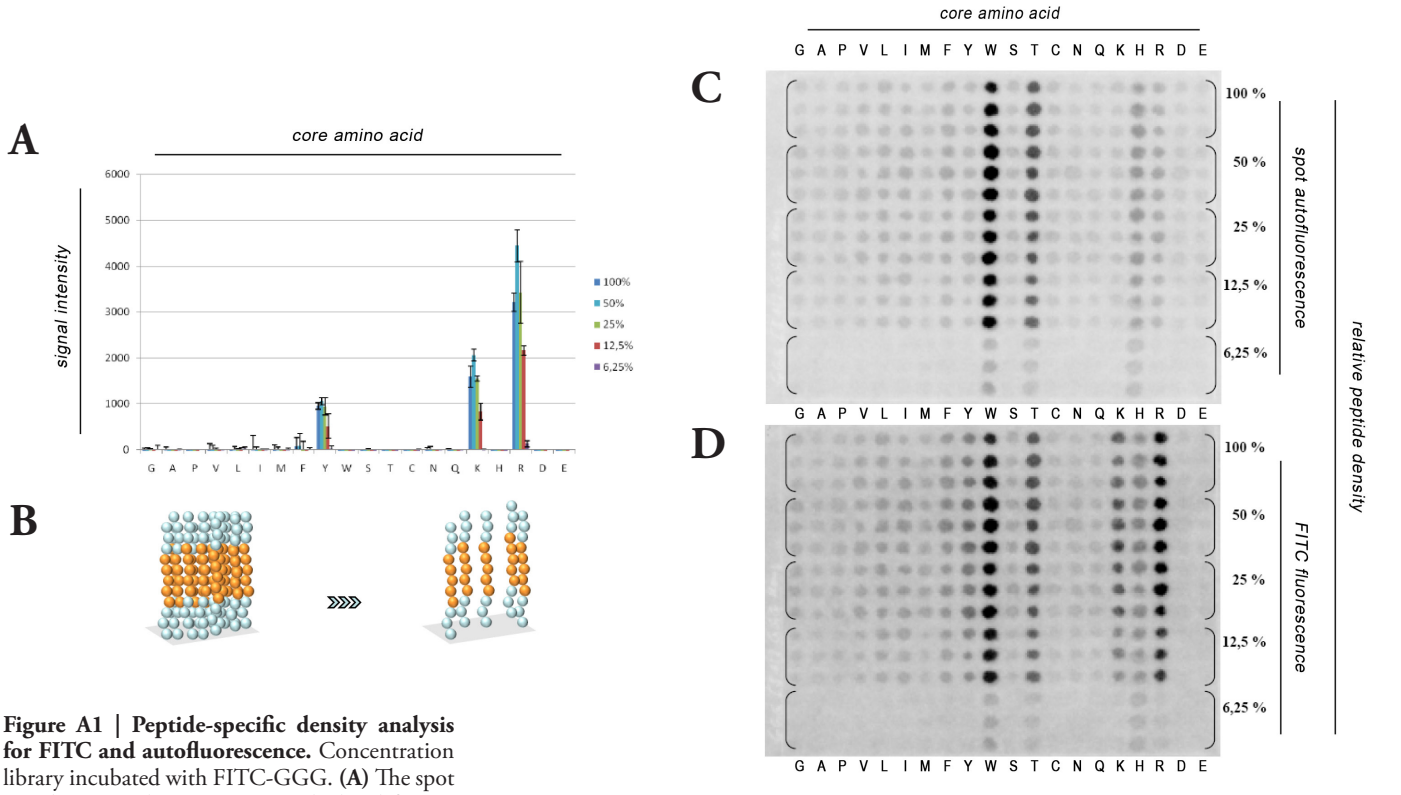


Figure A1 | Peptide-specific density analysis for FITC and autofluorescence. Concentration library incubated with FITC-GGG. **(A)** The spot signal measured at 520 nm is calculated from a circular region around the spot center detected in the image. SI is the background- (i.e., autofluorescence-) corrected calculated mean of three spots. Error bars represent the standard deviation of three spots. **(B)** Schematic overview. **(C)** Spot autofluorescence and **(D)** fluorescence at 520 nm. Each spot represents a cellulose membrane-bound peptide of the sequence GGG$[B]_5$GGG, where $[B]_5$ denotes five repeats of one of the 20 amino acids, and is repeated three times in different concentrations (100%, 50%, 25%, 12,5%, and 6,25%). Dark spots denote interaction. Contrast was adjusted to ensure better visibility of the spots.

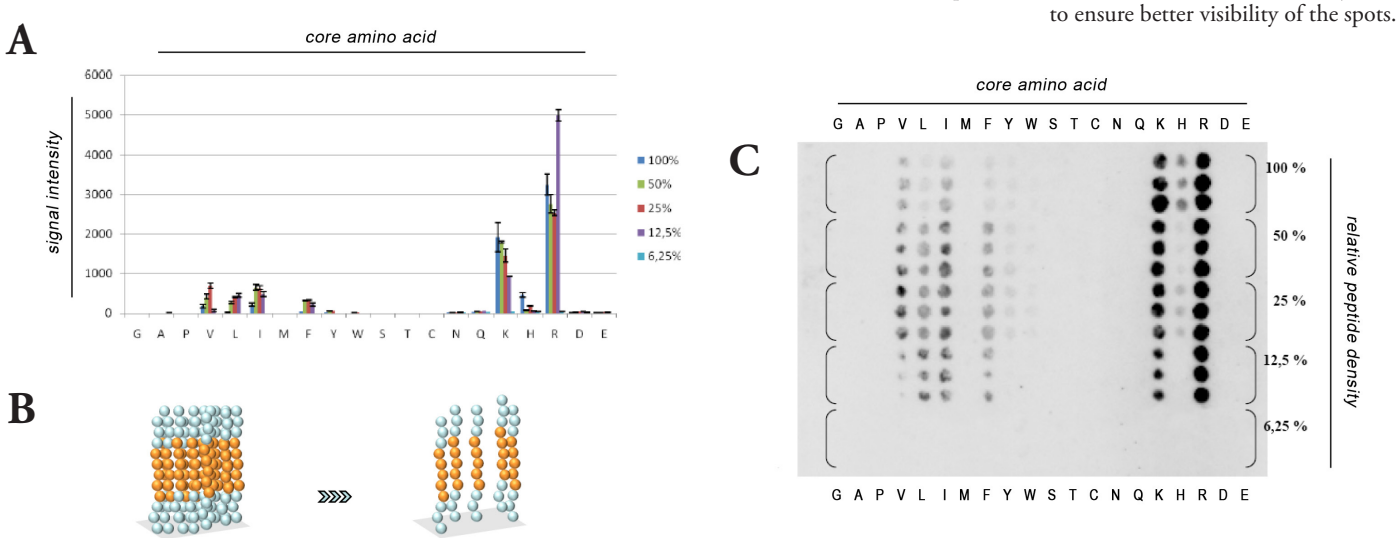


Figure A2 | Peptide-specific density analysis for biotin/streptavidin-POD. Concentration library incubated with biotin-GGG and detected with streptavidin-POD. **(A)** The spot signal measured by means of chemiluminescence is calculated from a circular region around the spot center detected in the image. SI is the background-corrected calculated mean of three spots. Error bars represent the standard deviation of three spots. **(B)** Schematic overview. **(C)** Each spot represents a cellulose membrane-bound peptide of the sequence GGG$[B]_5$GGG, where $[B]_5$ denotes five repeats of one of the 20 amino acids, and is repeated three times in different concentrations (100%, 50%, 25%, 12,5%, and 6,25%). Dark spots denote interaction. Contrast was adjusted to ensure better visibility of the spots.

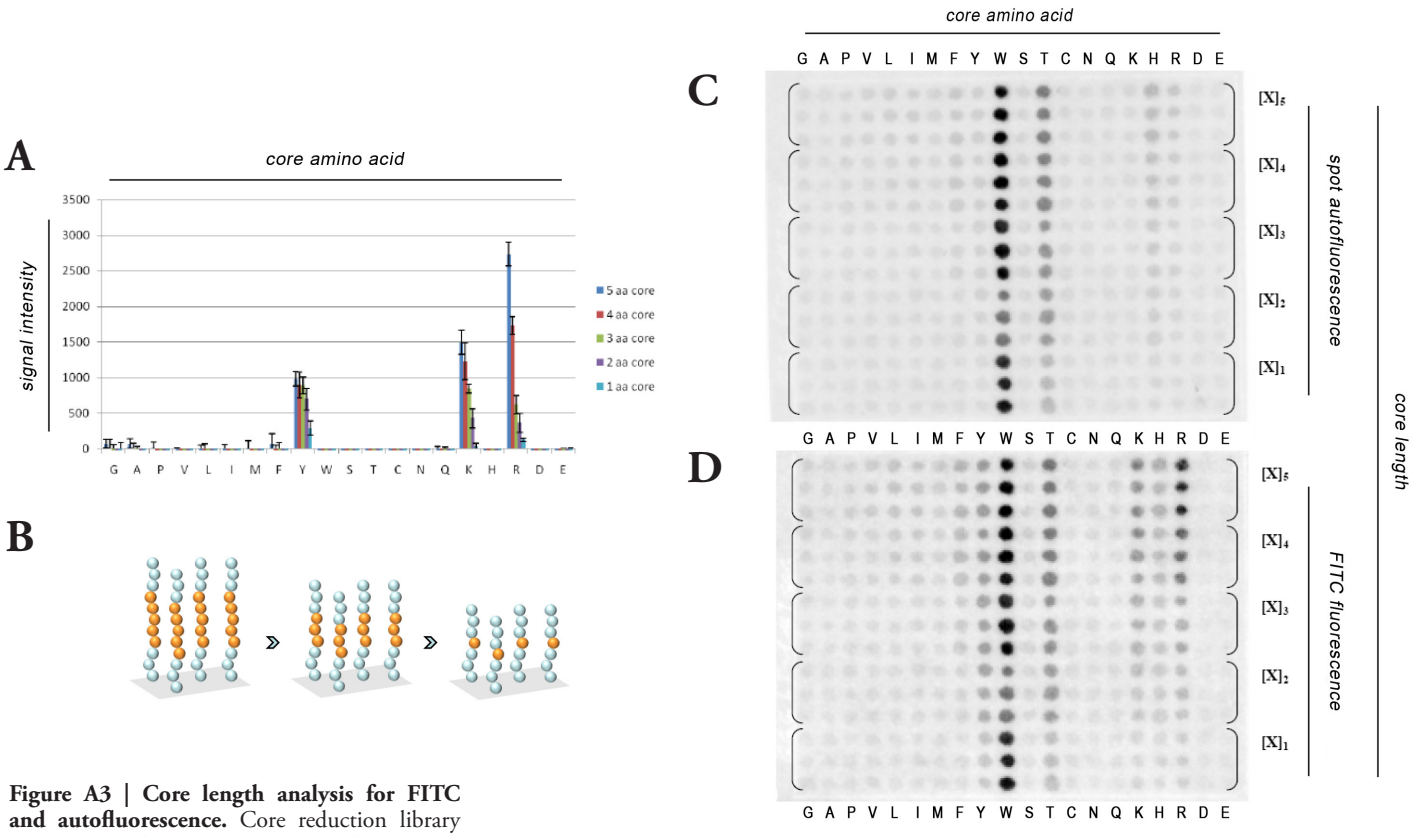
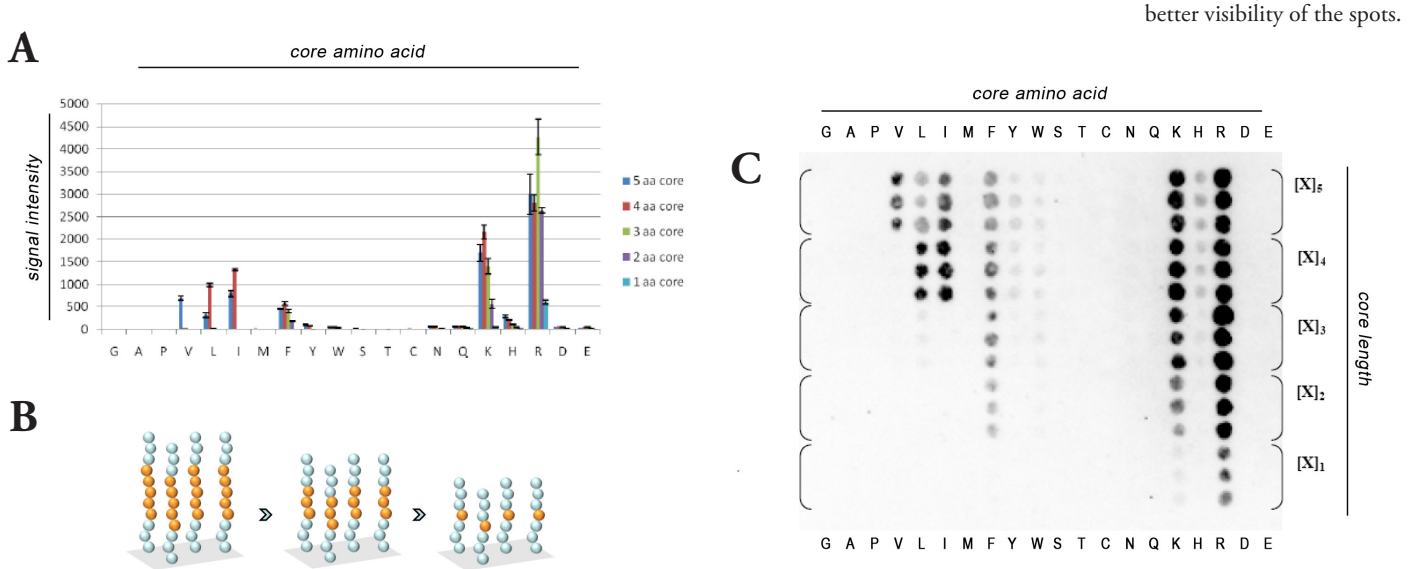


Figure A3 | Core length analysis for FITC and autofluorescence. Core reduction library incubated with FITC-GGG. **(A)** The spot signal measured at 520 nm is calculated from a circular region around the spot center detected in the image. SI is the background- (i.e., autofluorescence-) corrected calculated mean of three spots. Error bars represent the standard deviation of three spots. **(B)** Schematic overview. **(C)** Spot autofluorescence and **(D)** fluorescence at 520 nm. Each spot represents a cellulose membrane-bound peptide of the sequence GGG[B]$_5$GGG to GGG[B]$_1$GGG, where [B]$_{5–1}$ denotes 5–1 repeats of one of the 20 amino acids. Every spot is repeated three times. Dark spots denote interaction. Contrast was adjusted to ensure better visibility of the spots.

Figure A4 | Core length analysis for biotin/streptavidin-POD. Core reduction library incubated with biotin-GGG and detected with streptavidin-POD. **(A)** The spot signal measured by means of chemiluminescence is calculated from a circular region around the spot center detected in the image. SI is the background-corrected calculated mean of three spots. Error bars represent the standard deviation of three spots. **(B)** Schematic overview. **(C)** Each spot represents a cellulose membrane-bound peptide of the sequence GGG[B]$_5$GGG to GGG[B]$_1$GGG, where [B]$_{5–1}$ denotes 5–1 repeats of one of the 20 amino acids. Every spot is repeated three times. Dark spots denote interaction. Contrast was adjusted to ensure better visibility of the spots.



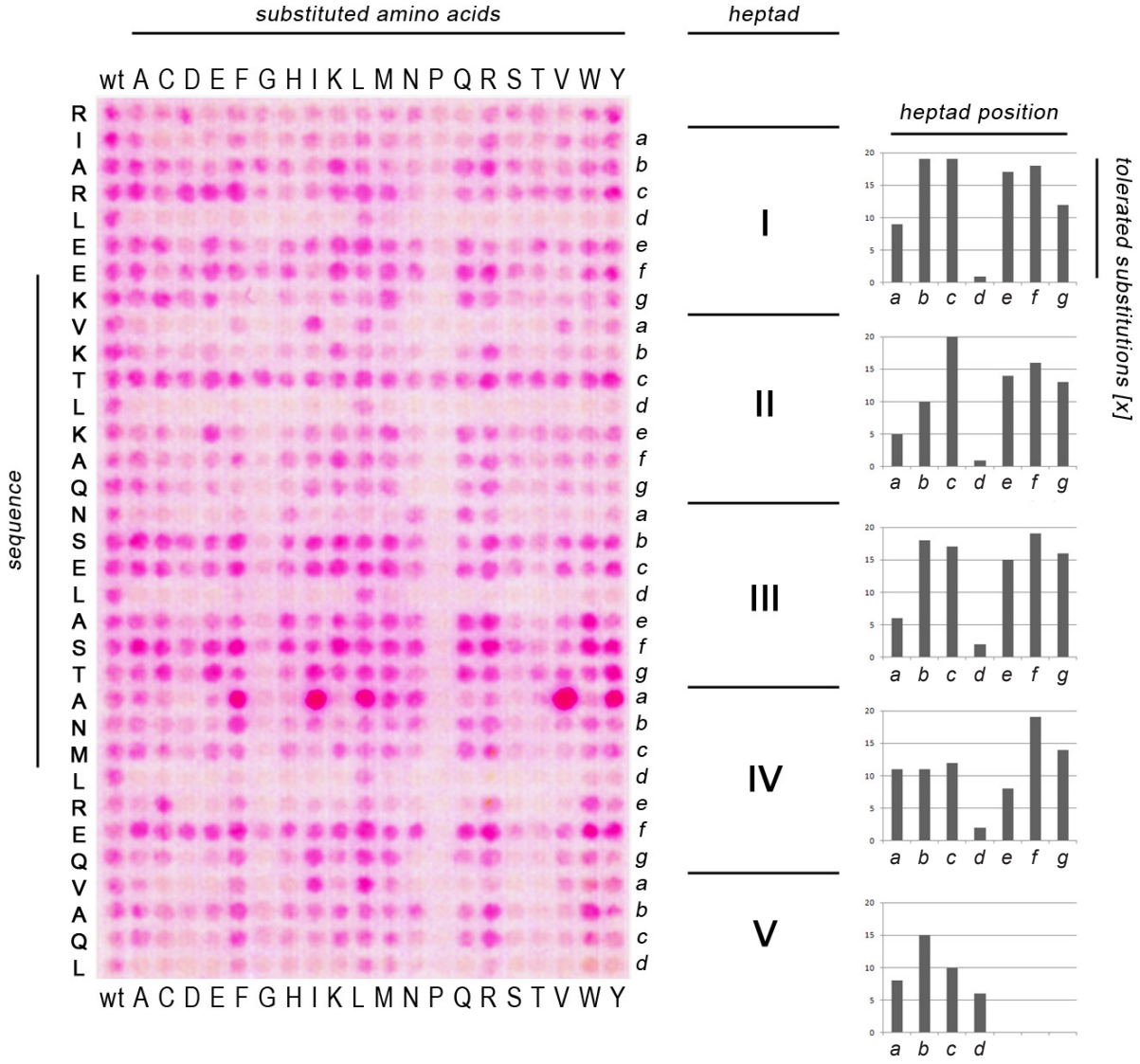


Figure A5 | Substitutional analysis of the homomeric c-Jun domain. (left) Red spots denote interactions between cellulose membrane-bound variants and a dye-labeled wildtype c-Jun sequence that was synthesized by standard solid-phase peptide synthesis and labeled with TAMRA at the N-terminus. Each spot corresponds to a variant in which one residue of the *wt* sequence given at the top was replaced by one of the 20 gene-encoded amino acids as specified on the left. Spots in the first row represent the *wt* sequence. **(right)** All spot signals of the array shown on the left were measured quantitatively, and successful replacements (countable binding spots) were determined. The quantity of tolerated substitutions is plotted against the positions in a given heptad.

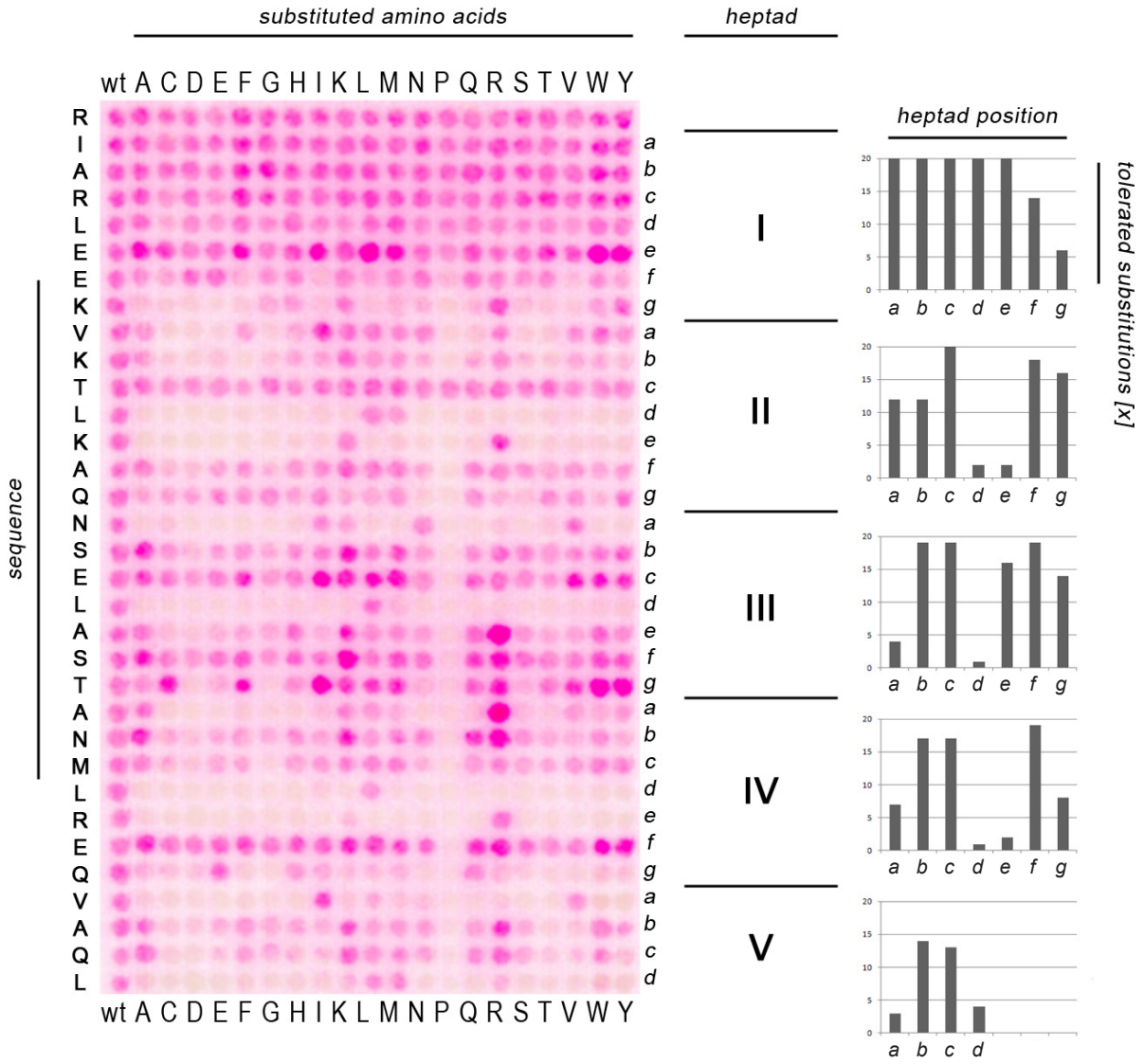


Figure A6 | Substitutional analysis of the heteromeric c-Jun/c-Fos domain. (left) Red spots denote interactions between cellulose membrane-bound variants of c-Jun and a dye-labeled wildtype c-Fos domain that was synthesized by standard solid-phase peptide synthesis and labeled with TAMRA at the N-terminus. Each spot corresponds to a variant in which one residue of the c-Jun$_{wt}$ sequence given at the top was replaced by one of the 20 gene-encoded amino acids as specified on the left. Spots in the first row represent the c-Jun *wt* sequence. **(right)** All spot signals of the array shown on the left were measured quantitatively, and successful replacements (countable binding spots) were determined. The quantity of tolerated substitutions is plotted against the positions inside a given heptad.

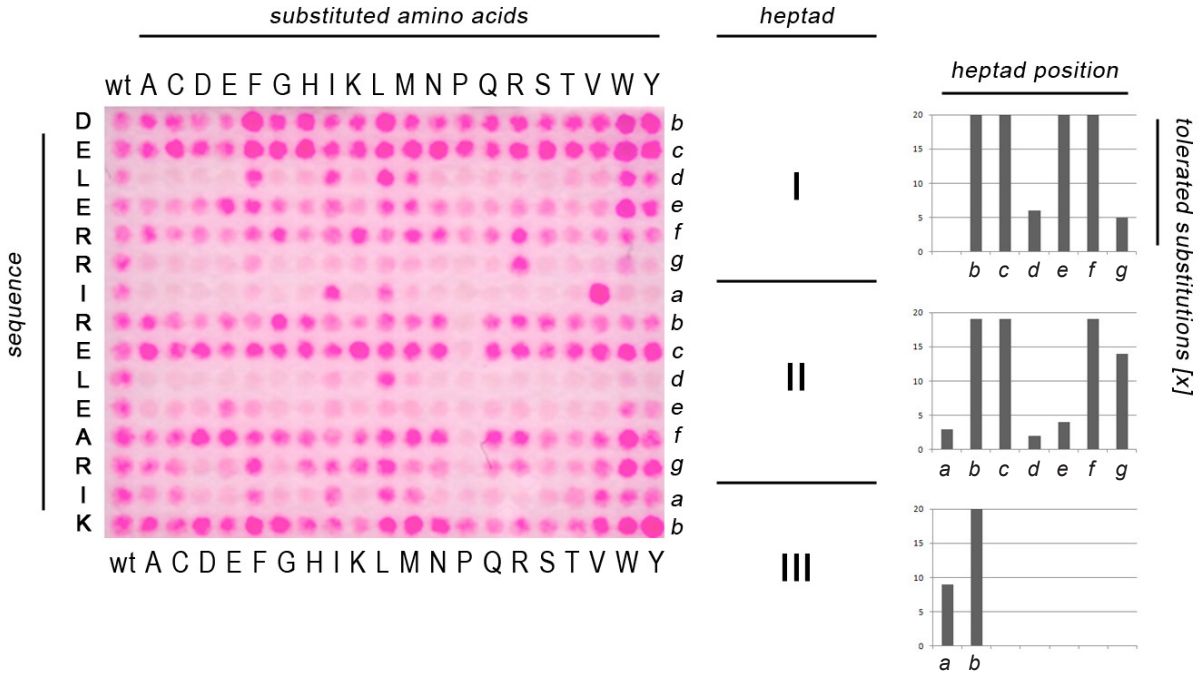


Figure A7 | Substitutional analysis of a designed short zipper domain. (left) Red spots denote interactions between cellulose membrane-bound variants of a dye-labeled short zipper domain [Burkhard *et al.*, 2000] that was synthesized by standard solid-phase peptide synthesis and labeled with TAMRA at the N-terminus. Each spot corresponds to a variant in which one residue of the short zipper sequence given at the top was replaced by one of the 20 gene-encoded amino acids as specified on the left. Spots in the first row represent the initial short zipper sequence. **(right)** All spot signals of the array shown on left were measured quantitatively, and successful replacements (countable binding spots) were determined. The quantity of tolerated substitutions is plotted against the positions inside a given heptad.

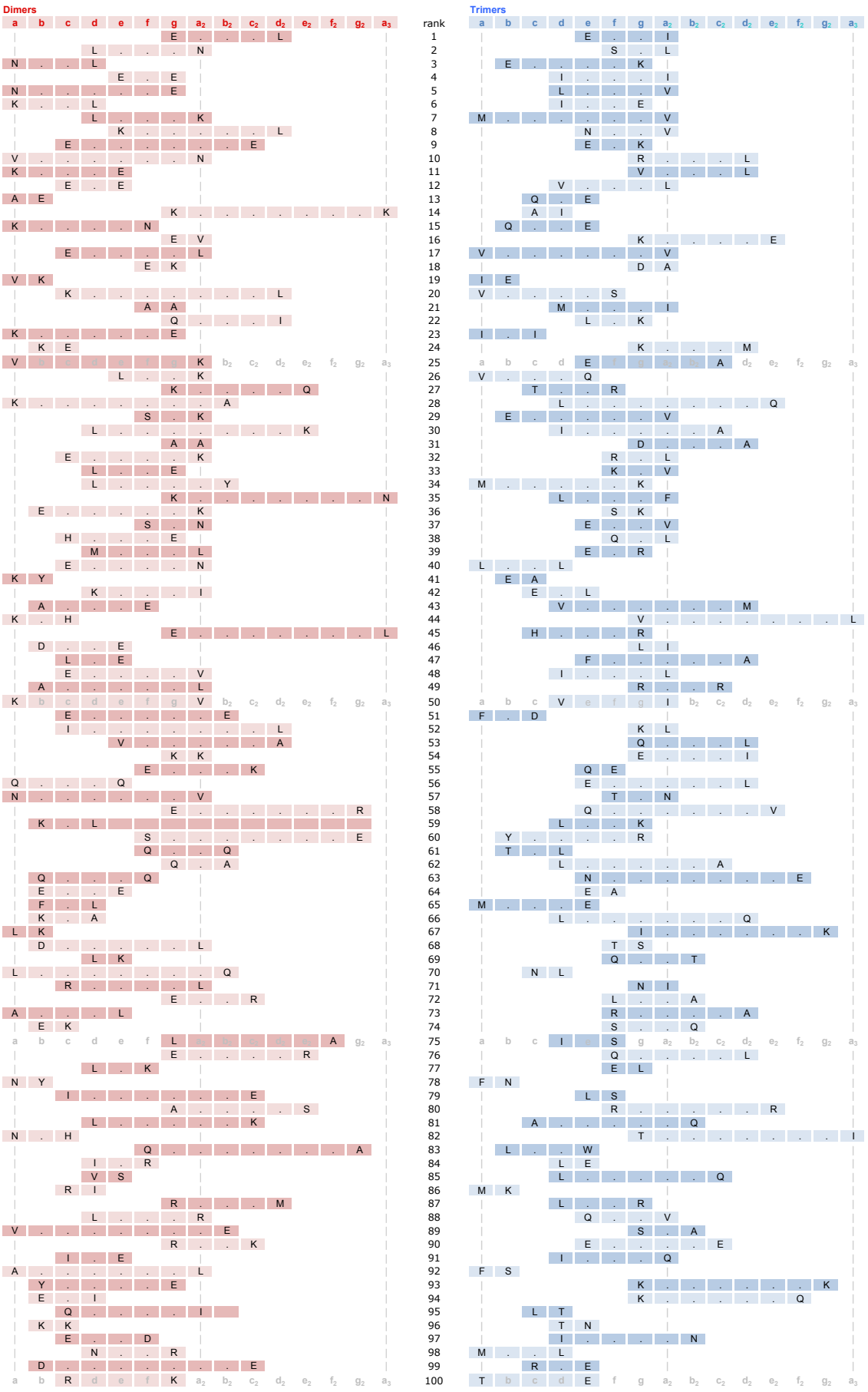


Figure A8 | List of the 100 most influential pairwise patterns. (left) The most influential patterns for dimers (highlighted in pink) and **(right)** trimers (highlighted in blue) are based on the PrOCoil model. The second dimer pattern L...N spanning columns d to a_2, for instance, describes a pattern with Leu at a *d* position, Asn at the next *a* position, and three arbitrary amino acids in between.

peptide	calc mass	mass found	purity
GGGGGGGGGGBB	787,73	809,125 [M + Na+]	85%
GGGAAAAAGGGBB	857,86	880,278 [M + Na+]	63%
GGGPPPPPGGGBB	988,05	988,25 [M]	34%
GGGVVVVVGGGBB	998,13	998,34 [M]	30%
GGGLLLLLGGGBB	1.068,26	1092,24 [M + Na^+]	75%
GGGIIIIIGGGBB	1.068,26	1090,29 [M + Na^+]	33%
GGGMMMMMGGGBB	1.158,45	1158,84 [M]	37%
GGGFFFFFGGGBB	1.238,35	1260,67 [M + Na^+]	65%
GGGYYYYYGGGBB	1.318,34	1318,69 [M]	30%
GGGWWWWWGGGBB	1.433,53	1433,56 [M]	31%
GGGSSSSSGGGBB	937,86	961,07 [M + Na^+]	57%
GGGTTTTTGGGBB	1.007,99	1008,08 [M]	55%
GGGCCCCCGGGBB	1.018,18	ND	ND
GGGNNNNNGGGBB	1.072,99	1096,68 [M + Na^+]	47%
GGGQQQQQGGGBB	1.143,12	1166,03 [M + Na^+]	39%
GGGKKKKKGGGBB	1.143,34	1167,11 [M + Na^+]	52%
GGGHHHHHGGGBB	1.188,17	1211,35 [M + Na^+]	66%
GGGRRRRRGGGBB	1.283,40	1282,7 [M]	45%
GGGDDDDDGGGBB	1.077,91	1077,99 [M]	25%
GGGEEEEEGGGBB	1.148,04	1171,63 [M + Na^+]	67%

Table A1 | MALDI-TOF and HPLC analysis of peptides synthesized on cellulose membranes. Peptides are linked to a β-alanine spacer (B). HPLC analysis was conducted using a linear solvent gradient (a: 0.05% TFA in water; b: 0.05% TFA in acetonitrile; gradient: 5–60% b over 30 min; UV detector at 214 nm; RP-18 column). α-cyanocinnamic acid was used as a matrix for MALDI-TOF MS analysis.

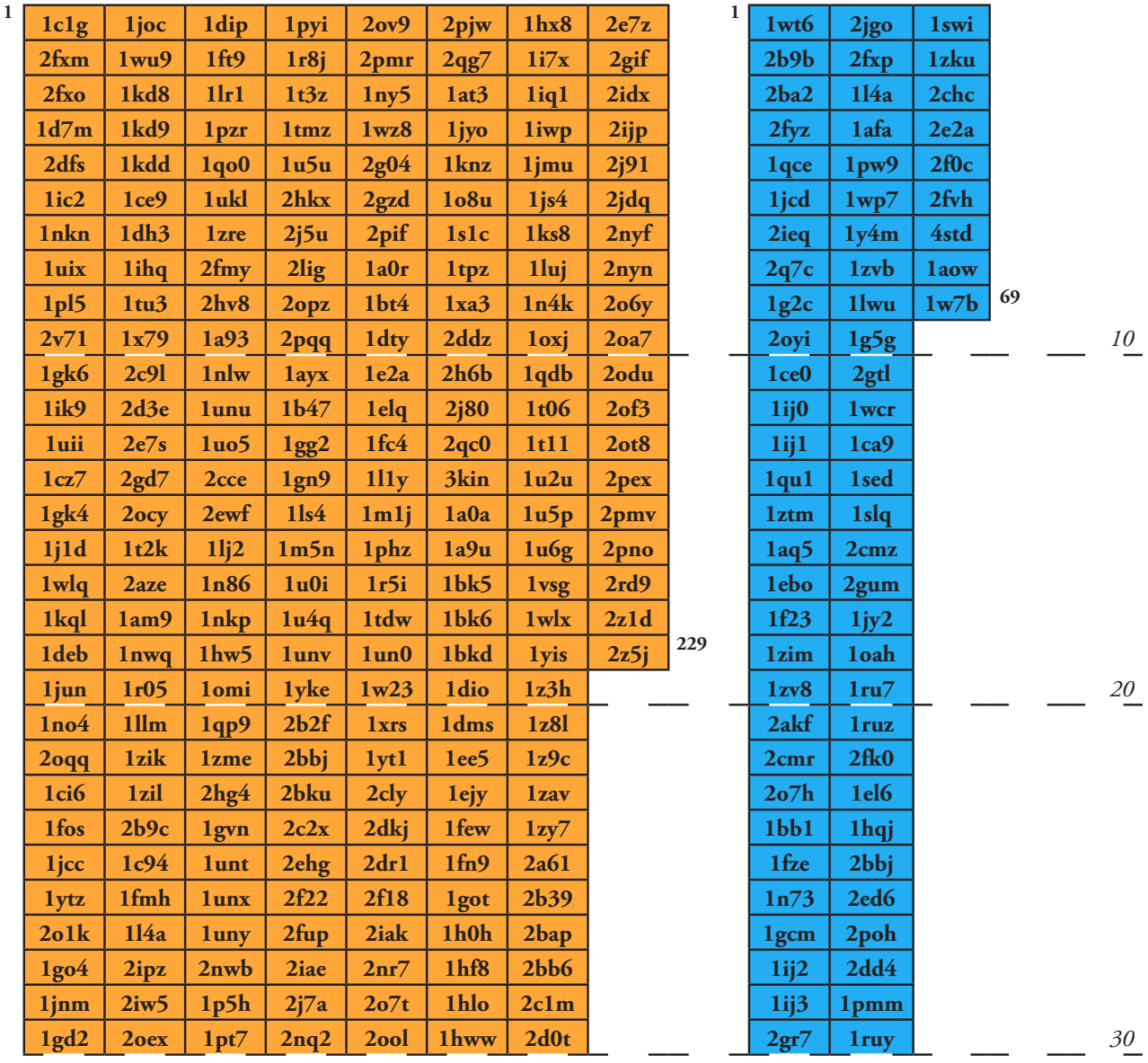


Table A2 | PDB identifier of the initial dataset of 385 dimeric and 92 trimeric structurally resolved coiled coil sequences. The program SOCKET [Walshaw et al., 2001] with a packing cutoff of 7.0 Å was used to scan the PDB for knobs-into-holes packing between helices. After further filtering steps, 229 dimeric (orange) and 69 trimeric (blue) entries were obtained. Note that entries may contain more than one coiled coil sequence due to hetero-oligomerization, discontinuous coiled coil regions (according to SOCKET), or multiple coiled coil regions within one sequence.

Publications

This work has been published in

peer-reviewed journals

Portwich, M., Keller, S., Strauss, H.M., **Mahrenholz, C.C**., Kretzschmar, I., Kramer, A., Volkmer, R., A Network of Coiled-Coil Associations Derived from Synthetic GCN4 Leucine Zipper Arrays. *Angew. Chem. Int. Ed.* **46**(10), 1654-1657 (2007)

and Portwich, M., Keller, S., Strauss, H.M., **Mahrenholz, C.C.**, Kretzschmar, I., Kramer, A., Volkmer, R., Ein mithilfe synthetischer GCN4-Leucinzipperarrays aufgedecktes Coiled-Coil-Assoziationsnetzwerk. *Angew. Chem.* **119**(10), 1682-1686 (2007)

Mahrenholz, C.C.*, Fidan, Z.*, Portwich, M., Volkmer, R., Analysis of coiled-coil associations by SPOT technology. *Chem. Today* **26**(2), 22-25 (2008)

Mahrenholz, C.C., Tapia, V., Stigler, R., Volkmer, R., A study to assess the cross-reactivity of cellulose membrane-bound peptides with detection systems: an analysis at the amino acid level. *J. Pept. Sci.* **16**, 297–302 (2010)

Mahrenholz, C.C.*, Abfalter, I.G.*, Bodenhofer, U., Volkmer, R., Hochreiter, S., Complex networks govern coiled coil oligomerization: Predicting and profiling by means of a machine learning approach. *submitted* (2010)

scientific meetings

Volkmer, R., Portwich, M., Strauss, H.M., **Mahrenholz, C.C.**, Kretzschmar, I., Keller, S., Analysis of coiled-coil associations by SPOT technology and biophysics. *8th German Peptide Symposium* (2007)

Mahrenholz, C.C., Promiskuitive Coiled-Coils und ihr pharmakologisches Potenzial. *6th Bionnale* (2008)

Abfalter, I.G.*, **Mahrenholz, C.C.***, Bodenhofer, U., Hochreiter, S., Analyzing Coiled Coil Proteins With Support Vector Machines to Design New Anti-Viral Drugs. *Intelligent Systems for Molecular Biology* (2008)

Mahrenholz, C.C., Tapia, V., Stigler, R.D., Volkmer, R., Peptides crossreacting with detection systems – Analysis at the amino acid level. *9th German Peptide Symposium* (2009)

Mahrenholz, C.C.*, Abfalter, I.G.*, Bodenhofer, U., Volkmer, R., Hochreiter, S., Defining distinctive rules for stoichiometry and recognition of coiled coil proteins using bioinformatical methods. *9th German Peptide Symposium* (2009)

Abfalter, I.G.*, **Mahrenholz, C.C.***, Bodenhofer, U., Hochreiter, S., New insights into coiled coil formation by means of support vector machines. *European Conference on Computational Biology and ISMB* (2009)

Mahrenholz, C.C., Portwich, M., Abfalter, I.G., Volkmer, R., Coiled coil formation: a multi-method approach to solve a glass bead game. *Protein Modules and Networks in Health and Desease* (2009)

scientific meetings (continued)

Volkmer, R., **Mahrenholz, C.C.**, Otte, L., Landgraf, C., Investigation of coiled coil domain based associations with synthetic peptide arrays: advantages and limitations *Protein Modules and Networks in Health and Desease* (2009)

Mahrenholz, C.C., Tapia, V., Stigler, R.D., Volkmer, R., A study to assess the cross-reactivity of cellulose membrane bound peptides with detection systems: An analysis at the amino acid level. *31st European Peptide Symposium* (2010)

Mahrenholz, C.C., Abfalter, I.G., Bodenhofer, U., Volkmer, R., Hochreiter, S., Complex networks govern coiled coil oligomerization: A multi-method approach. *31st European Peptide Symposium* (2010)

scientific meetings (with DOI publication)

Mahrenholz, C.C., Tapia, V., Stigler, R., Volkmer, R., Detection systems cross reacting with peptides – Analysis at the amino acid level. Available from *Nature Precedings* <http://dx.doi.org/10.1038/npre.2010.4448.1> (2010)

Bodenhofer, U., Kothmeier, A., Abfalter, I.G., **Mahrenholz, C.C.**, Hochreiter, S., Decoding Sequence Classification Models for Acquiring New Biological Insights. *18th International Conference on Intelligent Systems for Molecular Biology*. Available from *Nature Precedings* <http://dx.doi.org/10.1038/npre.2010.4708.1> (2010)

Mahrenholz, C.C.*, Abfalter, I.G.*, Bodenhofer, U., Volkmer, R., Hochreiter, S., PrOCoil - Advances in predicting two- and three-stranded coiled coils. *18th International Conference on Intelligent Systems for Molecular Biology*. Available from *Nature Precedings* <http://dx.doi.org/10.1038/npre.2010.4677.1> (2010)

patent

Mahrenholz, C.C., Portwich, M., Peptide für die Wechselwirkung mit alpha-helikalen Coiled-Coil-Strukturen und/oder Coiled-Coil-Sequenzen, davon abgeleitete Mittel und ihre Verwendung. *Patent PCT/DE2006/002295 (WO/2007/068240)*

other

peer-reviewed journal publication

Ay, B., Streitz, M., Boisguerin, P., **Mahrenholz, C.C.**, Schuck, S.D., Kern, F., Volkmer, R., Sorting and pooling strategy: a novel tool to map a virus proteome for CD8 T-cell epitopes. *Biopolymers* **88** (1), 64-75 (2007)

List of figures

List of tables

List of appendix figures and tables

Carsten Clemens Mahrenholz

Education

Professional Education

Military service

Academic studies

Scholarships and Awards

For reasons of data protection,
the curriculum vitae is not included in the online version.

Selected research interests

Activities

Advanced training

Achievements

For reasons of data protection,
the curriculum vitae is not included in the online version.