Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

APC Average Product Correction

DCA Direct Coupling Analysis

EC Evolutionary Coupling

MR Molecular Replacement

MSA Multiple Sequence Alignment

MX Macromolecular Crystallography

NOE Nuclear Overhauser Effect

SML Supervised Machine Learning

Chapter 1

Introduction

1.1 Macromolecular X-ray crystallography

The discovery of X-ray diffraction by crystals by Max van Laue [1, 2] marked the origins of modern crystallography. However, it was not until the work of William Lawrence Bragg and William Henry Bragg that X-ray scattering could be translated into atomic positions [3–5]. Since then, X-ray crystallography and the determination of atomic positions in organic and inorganic molecules of has come a long way and shaped the path for many 21st century discoveries. Amongst those ground-breaking discoveries are the earliest structural models of biological molecules including DNA [6], vitamin B12 [7], and the first protein structures [8–11]. These structure elucidations hallmarked the dawn of a new era in biological and biomedical research. At the time of writing, 124,551 structural models were determined by X-ray diffraction studies [12], and thus X-ray crystallography is a key method in biological research.

1.1.1 X-ray scattering

X-rays are high energy photons part of the electromagnetic spectrum with a wavelength 0.1-100Å [13]. X-rays can be described as packets of travelling electromagnetic waves, whose electric field vector interacts with the charged electrons of matter [13]. Such interaction, typically termed scattering, results in the diffraction of the incoming wave, which X-ray crystallography relies on.

In its simplest form, scattering of X-ray radiation can be explained in the scenario of exposure to a single free electron. The resulting scattering can be classed as elastic (Thomson scattering) or inelastic (Crompton scattering) [13]. The latter — scattering that results in a loss of energy of the emitting photon due to energy transfer onto the electron — does not contribute to discrete scattering, the type of scattering X-ray diffraction relies on. In comparison, Thomson scattering does not result in a loss of energy of the emitting photon. This has significant effects, the incoming photon emits with the same frequency causing the electron to oscillate identically further enhancing the signal.

If we expand the example to include all electrons in an atom and expose the atom to X-ray radiation, our theory needs to be slightly expanded. Given that one or more

electrons in an atom are not free but orbit around the atom's nucleus in a stable and defined manner, the distribution of these electrons around the nucleus determines the scattering of the incoming X-ray photons. The distribution of scattered photon waves is thus an overall representation of the probability distributions of each electron in the atom and is refered to as electron density $\rho(\boldsymbol{r})$. In X-ray scattering, it suffices to approximate the shape of the electron density to a sphere. If we now consider the emitting wave $\boldsymbol{s_1}$ of an X-ray photon scattered by any position $\boldsymbol{r}$ in the electron density of an atom, then the phase difference $\Delta\varphi$ to the incoming wave $\boldsymbol{s_0}$ can be described by Eq. 1.1 [13].

$$\Delta\varphi = 2\pi\left(\boldsymbol{s_1} - \boldsymbol{s_0}\right)\boldsymbol{r} = 2\pi \cdot \boldsymbol{Sr} \tag{1.1}$$

If more than one electron in an atom's electron density scatter the incoming X-ray wave, then the emitting partial waves can be described by the atomic scattering function f_s (Eq. 1.2), which describes the interference of all scattered waves [13]. The total scattering power of an atom is proportional to the number of electrons and element-specific with heavier atoms scattering more strongly. Given the approximation of a centrosymmetric electron density, the atomic scattering function is also symmetric.

$$f_s = \int_{\boldsymbol{r}}^{V(atoms)} \rho\left(\boldsymbol{r}\right) \cdot e^{2\pi i \boldsymbol{Sr}} \cdot d\boldsymbol{r} \tag{1.2}$$

With an enhanced understanding of X-ray scattering of electrons orbiting a single atom, it is important to consider X-ray scattering of adjacent atoms, such as it is typically found in molecules. If the electromagnetic wave of a X-ray photon excites all electrons of adjacent atoms, then the resulting partial waves result in constructive or destructive interference. Maximal interference can be obtained when all partial waves are in phase, and maximal destructive interference when out-of-phase. This leads to varying intensities of the emitting X-ray photon at different points in space. To obtain the overall scattering power F_s of all contributing atoms, Eq. 1.2 needs to be modified to include the sum over all atoms j as described in Eq. 1.3.

$$F_s = \sum_{j=1}^{atoms} f_{s,j}^0 \cdot e^{2\pi i \boldsymbol{S}\boldsymbol{r}_j} \tag{1.3}$$

If we now translate our hypothetical experiment into a crystal lattice then our understanding described in Eq. 1.3 needs to be expanded from a 1-dimensional distance vector $\boldsymbol{r}$ to the three dimensional lattice translation vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$. The Laue equations (Eq. 1.4) do exactly that and ultimately determine the positions of the diffraction peaks in 3-dimensional space.

$$\boldsymbol{S} \cdot \boldsymbol{a} = n_1, \qquad \boldsymbol{S} \cdot \boldsymbol{b} = n_2, \qquad \boldsymbol{S} \cdot \boldsymbol{c} = n_3 \tag{1.4}$$

$$n\lambda = 2d_{hkl}sin\theta \tag{1.5}$$

Such determination is possible through the findings made by Bragg and Bragg [3], who identified the relationship between the scattering vector $\boldsymbol{S}$ and the planes in the crystal lattice. Today, this relationship is defined by the Bragg equation (Eq. 1.5) [3], which allows us to interpret X-ray diffraction as reflections on discrete lattice planes, which relates the diffraction angle θ to the lattice spacing d_{hkl} (Fig. 1.1) [13]. For maximum diffraction n needs to be integer multiples to result in maximum constructive interference of wavelength λ.

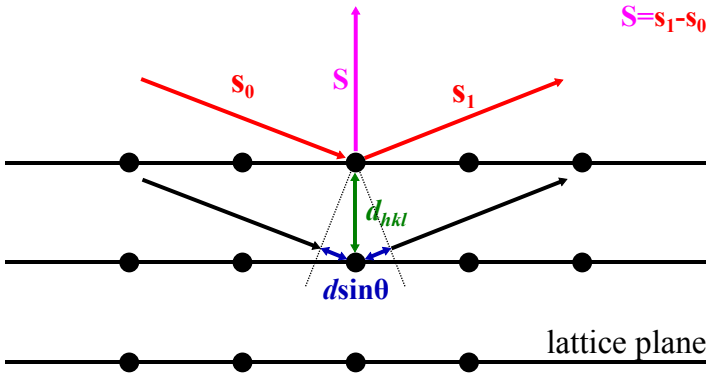

Figure 1.1: Schematic of Bragg scattering.

Lastly, if the hypothetical model is expanded to molecular crystals, then the total scattering from the unit cell is merely a summation of all molecular unit cell scattering contributions in the crystal. Mathematically, this results in Eq. 1.3 being generalised to Eq. 1.6 through the application of the Laue equations (Eq. 1.4) to express the scattering vector $\boldsymbol{Sr}_j$ as Miller indices of the reflection planes $\boldsymbol{hx}_j$.

$$F_h = \sum_{j=1}^{atoms} f_{s,j}^0 \cdot e^{2\pi i \boldsymbol{hx}_j} \tag{1.6}$$

The structure factor equation defines the scattering power from a crystal in a given reciprocal lattice direction $\boldsymbol{h}$. The scattering is enhanced by the number of repeating units of lattice translation vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$, and thus the overall scattering power is proportional to the number of unit cells in the crystal.

It should be noted that Eq. 1.6 is a simplification of the problem at hand. In reality, instrument and experimental corrections need to be applied to the structure factor equation. A correction factor for each experiment-dependent parameter needs to be applied to the structure factor equation. However, in the scope of this work the details of such correction factors do not need to be discussed.

$$\rho(x,y,z) = \frac{1}{V} \sum_{h=0}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} \boldsymbol{F}(hkl) \cdot e^{-2\pi i(hx+ky+lz)} \tag{1.7}$$

Since complex structure factors describe the molecular structure in the reciprocal space domain, the conversion to the real space domain in form of electron density is required. This can be conveniently done through the bijective Fourier transform, which allows to convert complex structure factors to electron density and vice versa without the loss of any information [13]. Thus, electron density can be obtained from the complex structure factors using Eq. 1.7. The normalisation factor $1/V$ provides the correct units for the electron density $\rho(x, y, z)$.

1.1.2 From crystal to structure

In X-ray crystallographic experiments, X-ray radiation is measured using light detectors. However, the measurement taken is incomplete. Light detectors only capture the intensity of the scattered X-ray photons but crucially lose the phase information. The latter is essential for atomic reconstruction of the molecule in the crystal, and thus needs to be obtained. In Macromolecular Crystallography (MX), experimentalists have a number of alternative techniques to compensate for the lost phase information.

Prior to the big advances in computing power and the successful elucidation of many protein structures, MX crystallographers primarily recovered the lost phase information through Direct Methods or Experimental Phasing [13]. Today, the most popular method to recovering the lost phase information is Molecular Replacement (MR) [14, 15]. In a MR search, a known structure ('search model') similar to the unknown is relocated in the unit cell until the solution with the best fit between calculated and observed diffraction data is obtained [13]. A 6-dimensional search, i.e. a simultaneous rotation and translation search, is possible [16–18], however computationally very expensive and less suitable for challenging cases. In comparison, most modern crystalloraphic applications opt for two distinct sub-searches, the rotation search to orient the search model within the unit cell followed by the translation search to locate it [13]. The benefits over a combined search include search-specific target functions that enable increased sensitivity and additional terms to compensate for imperfect data.

The most successful MR algorithms perform the rotation and translation searches using Patterson methods or Maximum Likelihood functions. Patterson methods — originally developed by Rossmann and Blow [19] — rely on the use of a map of vectors between the scattering atoms, which can be determined for the calculated and observed structure factor amplitudes. Patterson vectors can be sub-classed as intra- and inter-molecular vectors. A distinct separation of the observed vectors is impossible. However, inter-molecular vectors appear further away from the central peak of self-vector (vector from atom to itself) in the Patterson map [13]. The calculated Patterson vectors for the search model allow for a clearer distinction between the intra- and inter-molecular vectors. If the search model is placed in a large unit cell, then inter-molecular vectors must scale with the unit cell dimension [13]. Ultimately, using the intra-molecular Patterson vectors, the search probe can be oriented against the experimentally determined Patterson vectors. Similarly, the inter-molecular vectors can be used to identify the correct translation of the search probe. Patterson methods are very sensitive to small orientation errors of the search probe [13]. Thus, orientations with the highest vector peak overlaps are trialed in the subsequence translation search.

In comparison to the Patterson methods, Maximum Likelihood methods do not rely on inter-atomic vectors in Patterson maps. Instead, Maximum Likelihood methods make use of Bayes' theorem [20] to compare calculated structure factors and observed structure factor amplitudes directly [18]. Bayes' theorem in crystallographic Maximum Likelihood methods is applied to compute the likelihood that an experimental value is observed given the current search model. The maximal likelihood indicates the best search model given the observed experimental data. Since the search model likelihood term is the product of many individual probabilities, which are difficult to represent computationally due to floating point representations, the log of the likelihood is commonly used [13]. The major advantage of Maximum likelihood methods over Patterson methods centres on the more realist target functions, which consider errors and incompleteness of the search model, applies bulk solvent correction and conducts multi-model searches [18]. The latter is of particularl relevance since the Maximum likelihood rotation function can thus consider already placed search model probes in a fixed position whilst trialling additional ones [21], which proves to be a major advantage over Patterson methods. Furthermore, likelihood

target functions consider the structural variance of multiple superposed models in an ensemble search model, which is used to weight structure factors at the various positions to improve the overall likelihood term [18].

The initial electron density map after MR is almost always inaccurate because of the search model-based phases. Inaccuracies arise from experimental errors, model incompleteness, low signal-to-noise or model bias. Thus, approaches for improving the phases used to calculated the initial electron density map have been developed and are routinely applied in MX. Density modification describes a set of methods that improve the obtained electron density typically by applying statistical corrections to electron density distributions. These corrections are based on prior knowledge or assumptions of the physical properties of macromolecular structures [13]. This process can transform initially poor or uninterpretable initial electron density maps to high quality ones. Three pre-dominant density modification approaches exist: solvent flattening, histogram matching and the "sphere-of-influence" method. Solvent flattening is an approach first proposed by Wang [22], which exploits the fact that solvent regions in protein crystals are disordered, and thus differ in electron density volume from macromolecule-containing regions. If solvent electron density is set to a constant, then it is essentially flattened which will result in improved structure factors with improved phases and thus improved electron density. Histogram matching [23] exploits the defined characteristics of an electron density distribution determined from sets of proteins at the same resolution, irrespectivce of individual structural details. The electron density distribution for noisy maps are Gaussian-shaped. In contrast, the electron density distribution of a feature-defined map is positively skewed. The "sphere-of-influence" method was introduced by Sheldrick [24] and classifies solvent and protein electron density by observing its variance across the shell surface of a 2.42Å sphere (dominant 1-3 atom distance in macromolecular structures). If the sphere is positioned in the disordered solvent region typically found in intermolecular channels, the density variance will be low. Thus, this approach allows to smoothen solvent-containing regions of the electron density [24]. Independent of the density modification strategy applied, it is important to understand that improvements to the electron density map anywhere lead to improvements everywhere by transferral of information from one part of the map to another [25].

A second approach to improving the initial electron density is termed Refinement. Iteratively, the placed search model is optimised to better describe the experimentally observed data. This optimisation problem is typically broken down into three main steps: the definition of the model parameters, the scoring function and the optimisation method. The model parameters describe the crystal and its content and can be subdivided into atomic and non-atomic model parameters [26]. These parameters combined are used to score the current model. The scoring function relates the experimental data to the model parameters. The scoring function contains two primary terms, the refinement data target and an *a priori* knowledge term. The former defines a target function that assesses the similarity between calculated and experimental structure factors. The target function is commonly a Maximum Likelihood-based function that considers missing or incomplete data [26, 27]. The *a priori* knowledge term in the scoring function defines the properties of a good model by including stereochemical property terms. Lastly, optimisation methods provide tools to vary the model parameters to better fit the experimental data. Different optimisation techniques can be used depending on the severity of model parameter alteration, which generally depend on the entrapment of states in local energy minima. The three steps combined form a macrocycle that iteratively modifies the model to optimise its fit to the experimental data. This ultimately improves both the electron density map interpretability and model quality. MX refinement can be performed in structure-factor-based reciprocal space and electron-density-based real space [26]. A combination allows global and local refinement strategies and enables grid-like searches to optimise the model parameters until convergence.

Once initial phase information is improved through refinement and density modification, attempts can be made to build atomic model coordinates into the electron density map. This process is typically coupled with refinement or density modification to iteratively improve the quality of the partially built model and the electron density map [13]. A small number of distinct algorithms are currently used to automatically build atomic coordinates into electron density: main-chain autotracing [28], fitting pseudo-atoms into electron density [29], or fitting reference coordinates with similar electron density maps [30, 31]. In essence, all algorithms attempt to maximise the number of correctly identified and placed atomic coordinates into available electron density. Whilst autotracing solely

builds main-chain peptides, the other two approaches rely on sequence information to also build side-chains. Independent of the complexity of the model building task, the higher the resolution and the more complete the initial starting model, the less ambiguous and challenging this overall task becomes [13].

1.1.3 Unconventional Molecular Replacement

The process of macromolecular structure determination via conventional MR has been outlined previously. Search models are typically derived from structural homologs identified by sequence similarity to the crystallised target [13]. However, with decreasing sequence similarity between homologs, it becomes more challenging to identify structural templates suitable for MR. Furthermore, experimental phasing approaches to circumvent the absence of MR templates can be expensive, unsuccessful and very challenging for certain protein targets, and thus remain unfeasable to pursue at times. Under such circumstances, alternative approaches are required, which are referred to as "unconventional" MR approaches from here onwards. The unconventional MR approach most relevant to the work presented in this thesis utilises the 3-dimensional structure prediction of a protein target starting from its sequence [32–34].

1.2 *Ab initio* protein structure prediction

The folding of protein structures is commonly described by the folding funnel hypothesis [35]. It assumes that the native state of a protein fold corresponds to its global minimum free energy state along its energy surface (Fig. 1.2) [36]. *In silico* protein folding experiments attempt to find this lowest-free-energy state of the protein fold; however, to unambiguously identify it sampling of all polypeptide chain conformations is necessary. In theory, sampling of all conformations for a 100-residue protein takes in the order of approximately 10^{52} years (10^{7} configurations with 10^{-11} seconds per configuration), yet in practice an equivalent polypeptide chain would fold in milliseconds to seconds [37, 38]. This paradox — termed the Levinthal paradox [37] — created the basis for the folding funnel hypothesis.

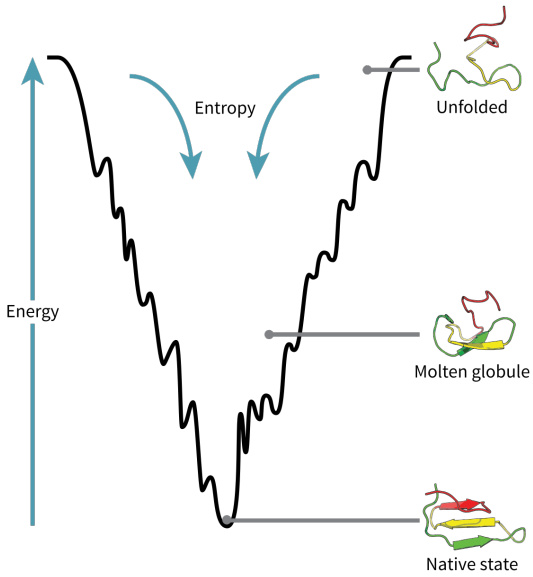

Figure 1.2: Schematic of the folding funnel hypothesis [35]. Diagram produced by Wikipedia [39] contributors.

In *ab initio* protein structure prediction, the tertiary structure of a protein is predicted using its primary structure alone. This problem is in its nature identical to finding the lowest-energy state along the protein's energy landscape. However, in an attempt to avoid the Levinthal paradox, different knowledge- and physics-based energy functions coupled with a variety of conformational-search sampling algorithms are employed [40].

Physics-based energy functions use physiochemical force fields typically coupled with Molecular Dynamics simulations to sample the folding trajectory of a protein sequence (true physics-based approaches are computationally intractable because quantum mechanics models would need to be used). Force fields describe parameter sets used to calculate energy potentials for a system of atoms in a simulation run, and include potentials such as van der Waals and electrostatic interactions [40]. In the context of *ab initio* protein structure prediction, pure physics-based approaches are often less favourable, because the computational complexity to find the lowest free-energy state of a large protein structure remains intractable without the use of supercomputers.

Knowledge-based energy functions rely on empirical energy terms derived from statistics and regularities of experimentally determined structures [40]. These energy terms can be subdivided into two types, the generic or sequence-independent terms and amino-acid or sequence-dependent terms [41]. The former include terms to describe the backbone

hydrogen-bonds and local backbone stiffness of a polypeptide chain. The latter describes terms such as pairwise residue contact potential, distance-dependent atomic contact potential, and secondary structure propensities. However, predicting local or global tertiary structure of a protein sequence using empirical energy terms alone is very difficult. Subtle differences in the local and global environment of a primary structure alongside the subtle differences in primary structures leading to common secondary structure features are very difficult to reproduce in a modelling scenario. Thus, knowledge-based energy functions are often coupled with the assembly of fragments extracted from other protein structures to predict the unknown tertiary structure of the target sequence [40].

The most successful *ab initio* structure prediction protocols use knowledge-based and physics-based energy functions combined with fragment-assembly-based conformational searches to find the lowest free-energy state [42–46]. Structural fragments of varying lengths (typically 3-20 residues) are extracted from existing protein structures [47–54]. These fragments are used in a Monte-Carlo simulation to search the conformational space of the polypeptide chain to search for low free-enery states [55]. The insertion of overlapping fragments results in the replacement of torsion angles either at random positions or sequentially from pre-defined starting position (such as N- or C-termini), and each move is scored against the Metropolis criterion [55] consisting of knowledge-based and physics-based terms. If a fragment passed the Metropolis criterion, its torsion angles are accepted and integrated in the polypeptide chain for the next fragment-insertion iteration. This process is repeated until convergence of the decoy, i.e. no lower free-energy state can be found. In all routines, these steps are independently repeated thousands of times to create a pool of decoys.

In order to identify the correct fold amongst the thousands of generated decoys, clustering approaches are commonly in combination with *ab initio* protocols. Shortle et al. [56] identified that the most-similar decoy to the native structure is most often the centroid (decoy with most neighbours in the cluster) of the largest cluster. Further studies showed that the selection of those centroid decoys helps to identify the most native-like folds amongst the many thousands generated [57–59]. Some protocols use clustering as an intermediate or final step to identify decoys for which it will perform more computationally demanding all-atom refinement [58] or other decoy hybridisation [43, 60, 61] approaches

to further approach the native-like fold [62].

Despite active research in *ab initio* protein structure prediction over decades, all approaches cannot reliable predict high-resolution structures for anything but small globular folds [58, 63–65]. The major issue arises from the sampling of the conformation space since incorrect local changes influence the global structure. Furthermore, β-sheets are inheritely difficult to predict given that β-strands in fragment-based approaches are inserted one at a time yet rely on the hydrogen bond network typically found in β-sheets to reduce the overall energy of the decoy. To address this issue, Lange et al. [66], Raman et al. [67], and Göbl et al. [68] started to use Nuclear Overhauser Effect (NOE) data as residue-residue distance restraints to reduce the sampling space of conformations, which enabled high-resolution prediction of tertiary structure for longer protein peptides. Nevertheless, only the use of residue-residue contact information as proxy for spatial proximity of amino acid pairs enabled accurate *ab initio* structure prediction for longer polypeptide chains (e.g., [45, 46, 69–75]).

1.3 Residue-residue contact prediction

The use of residue-residue contact information to reduce the conformational search space in protein structure prediction relies on the identification of amino acids in close spatial proximity. Today, such identification can be detected from sequence information alone by either Direct Coupling Analysis (DCA) or Supervised Machine Learning (SML) algorithms.

1.3.1 Direct Coupling Analysis

Direct Coupling Analysis relies on the use of protein sequence information to identify coordinated changes of amino acids in sequences of a protein family (Fig. 1.3). These coordinated changes are caused by evolutionary pressure to maintain residue interactions important for protein structure and function. However, original attempts to detect covariation signal from sequences in a protein family were unsuccessful for many years [76–79]. The applied (local) statistical model suffered from numerous drawbacks, including

the loss of covariation signal due to phylogenetic dependencies, limited availability of sequence data, and the potentially false assumption that truly coevolved residues are in close proximity [80–82]. Implementations of the local statistical model used raw covariation frequencies between pairs of positions in the sequence alignment. This further poses issues since successful distinction between "direct" casual (A-B and B-C) and "indirect" transitive (A-C) correlations is essential for successful protein structure prediction.

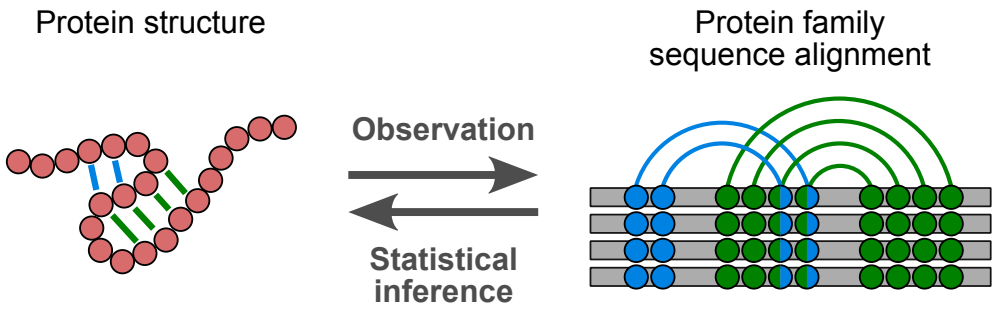

Figure 1.3: Schematic of inference of covariance signal originating from evolutionary pressure in protein tertiary structures and encoded in its family's sequence alignment (adapted from [83]).

Lapedes et al. [81] proposed the use of a global statistical model to infer correlations of residue pairs to circument the main problem of decoupling causal and transitive correlations. However, it was not until a decade later before first implementations of the global statistical model surfaced to successfully disentangle these types of correlations [69, 84–91]. Global statistical models achieve successul disentanglement by inferring a probabilistic description of the sequence alignment that best explains observed correlations using underlying causal couplings between positions [92]. Such couplings can be inferred by maximising the likelihood of observing the sequences in the alignment under the maximum entropy probability model. In other words, by considering all amino acid pair positions simultaneously, causal and transitive couplings can be successfully disentangled [89].

The pairwise probabilistic model $P(\boldsymbol{\sigma})$ of the amino acid sequence $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ of length N is defined in Eq. 1.8, which contains the amino acid configuration constraints σ_i and σ_j at positions i and j, the single-site conservation bias term h_i, and co-conservation term J_{ij} between position pairs i, j.

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1 \leqslant i < j \leqslant N} J_{ij}(\sigma_i, \sigma_j)\right) \tag{1.8}$$

The partition function Z (Eq. 1.9) acts as normalising constant, and additionally has the property to maximise the entropy in the probabilistic model. However, the computation of Z is intractable for the feature space found in DCA since the number of summations in Z exponentially increases with N for all 20 amino acid configurations. Thus, approximations of Z are typically used, which were shown lead to precise covariance predictions [89].

$$Z = \sum_{\boldsymbol{\sigma}} \exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1 \leqslant i < j \leqslant N} J_{ij}(\sigma_i, \sigma_j)\right) \tag{1.9}$$

Over the last decade, numerous approximations for the parameter inference of $P(\boldsymbol{\sigma})$ have been implemented, which include gradient ascent with Monte Carlo sampling [82], message passing [84], mean-field [69, 87, 88, 93], and pseudolikelihood maximisation [86, 89–91, 94]. However, it is the latter that has proven to be most successful, and thus it is commonly used in most applications. In pseudolikehood maximisation DCA approaches, the full likelihood for each sequence position i in $\boldsymbol{\sigma}$ across all sequences in the alignment is approximated by a product of conditional likelihoods (Eq. 1.10) [92].

$$\mathcal{L}(\mathbf{h}, \mathbf{J}) = \prod_{\sigma \in \Sigma} P(\sigma | \mathbf{h}, \mathbf{J}) \approx \prod_{i=1}^{N} P(\sigma_i | \sigma \backslash \sigma_i, \mathbf{h}, \mathbf{J}) \tag{1.10}$$

Equation 1.10 describes the conditional probability of observing amino acid (σ_i) in position i given all other amino acids ($\sigma \backslash \sigma_i$) in $\boldsymbol{\sigma}$. This leads to the cancellation of the partition function Z, and instead normalises locally over all possible 20 amino acid configurations at each site i. The parameters $\mathbf{h}$ and $\mathbf{J}$, which minimise Eq. 1.10, are identified using iterative optimisation algorithms [92]. Typically, regularisation terms are also added to Eq. 1.10 to avoid overfitting of the input data [92].

The positional constraint matrices J_{ij} for all amino acid (k) pairs across all combinations of σ_i and σ_j in $\boldsymbol{\sigma}$ need be summarised to a coupling score between σ_i and σ_j. The Frobenius norm is the preferred summary statistic (Eq. 1.12), and applied to a row- and column-means-centered coupling matrix J'_{ij} (Eq. 1.11). Furthermore, Average Product Correction (APC) is applied to remove background coupling that arises due to noise from phylogentic relationships between sequences to provide the final evolutionary coupling Evolutionary Coupling (EC) score (Eq. 1.13) [88–91, 95].

$$J'_{ij} = J_{ij}(k,l) - J_{ij}(\cdot,l) - J_{ij}(k,\cdot) + J_{ij}(\cdot,\cdot) \tag{1.11}$$

$$FN(i,j) = \sqrt{\sum_k \sum_l J'_{ij}(k,l)^2} \tag{1.12}$$

$$EC(i,j) = FN(i,j) - \frac{FN(i,\cdot)FN(\cdot,j)}{FN(\cdot,\cdot)} \tag{1.13}$$

Despite the great precision achievable by DCA algorithms, such algorithms suffer from one major drawback. All covariance-based algorithms rely on sufficiently large and diverse Multiple Sequence Alignment (MSA)s. Although the minimum number of sequences required per MSA might be target- and algorithm-dependent, early works suggested a minimum required of > 1000 sequence homologs [88, 96, 97]. Simultaneously, Marks et al. [69] and Kamisetty et al. [90] recommended a more sequence-specific length-dependent factor, whereby the sequence count in the alignment should exceed at least five times protein length for precise predictions. Whilst those earlier suggestions permit crude estimations of the likelihood of obtaining precise contact predictions, researchers realised that highly redundant MSAs could surpass such a threshold yet not provide enough diversity typically required for covariance-signal detection. Thus, the measure of *alignment depth* (also termed *number of effective sequences*) was introduced to capture both the sequence count and diversity in a given alignment [87, 98–100]. Although target- and

algorithm-dependent threshold persist, a minimum of 100-200 effective sequences are typically required [99, 100]. Furthermore, individual sequence weights used to calculate the alignment depth are widely used in covariance-based algorithms to reweight individual sequences to reduce the phylogenetic effect of non-independently evolved sequences in the MSA [89].

1.3.2 Supervised Machine Learning

Unlike DCA approaches, Supervised Machine Learning algorithms do not rely on the availability of homologous sequences to predict residue-residue contacts. Instead, SML models are trained on a variety of sequence-dependent and sequence-independent features to infer contacting residue pairs [101–106]. Broadly speaking, such SML algorithms rely on the analysis of sequence-based features, such as secondary structure, and sequence profiles. SML algorithms suffer from a similar inability to distinguish between residue pairs that form direct and indirect contact pairs, similar to earlier implementations of covariance-based methods. However, pure SML-based algorithms are not relevant to the work described in this thesis, and thus not further discussed. It is worth noting though that covariance-based algorithms outperform pure SML algorithms for protein families with many homologous sequences. However, SML algorithms do outperform DCA algorithms for families with fewer homologous sequences [99, 106, 107].

1.3.3 Contact metapredictors

The most recent approaches in residue-residue contact prediction use combinatorial approaches to exploit information from DCA and SML approaches. Metapredictors commonly use SML approaches as priors [71] or posteriors [75, 99, 100, 108–110] in addition to DCA algorithms. Furthermore, metapredictors use multiple input MSAs and/or DCA algorithms to further enhance the prediction precision. In most cases, metapredictors outperform their individual approaches and improvements are most noticable for targets with lower alignment depths [75, 111, 112].

Bibliography

[1] W Friedrich, P Knipping, M Laue, *Ann. Phys.* **1913**, *346*, 971–988.

[2] M Laue, *Ann. Phys.* **1913**, *346*, 989–1002.

[3] W. H. Bragg, W. L. Bragg, *Proceedings of the Royal Society A: Mathematical Physical and Engineering Sciences* **July 1913**, *88*, 428–438.

[4] W. L. Bragg, *Scientia* **1929**, *23*, 153.

[5] W. L. Bragg, *Nature* **Dec. 1912**, *90*, 410.

[6] J. D. Watson, F. H. C. Crick, Others, *Nature* **1953**, *171*, 737–738.

[7] D. C. Hodgkin, J Kamper, M Mackay, J Pickworth, K. N. Trueblood, J. G. White, en, *Nature* **July 1956**, *178*, 64–66.

[8] T. L. Blundell, J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. C. Hodgkin, D. A. Mercola, M Vijayan, en, *Nature* **June 1971**, *231*, 506–511.

[9] C. C. Blake, D. F. Koenig, G. A. Mair, A. C. North, D. C. Phillips, V. R. Sarma, en, *Nature* **May 1965**, *206*, 757–761.

[10] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H Muirhead, G Will, A. C. North, en, *Nature* **Feb. 1960**, *185*, 416–422.

[11] J. C. Kendrew, G Bodo, H. M. Dintzis, R. G. Parrish, H Wyckoff, D. C. Phillips, en, *Nature* **Mar. 1958**, *181*, 662–666.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **Jan. 2000**, *28*, 235–242.

[13] B. Rupp, *Biomolecular crystallography : principles, practice, and application to structural biology*, English, Garland Science, New York, **2010**.

[14] M. G. Rossmann, *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1360–1366.

[15] M. G. Rossmann, en, *Acta Crystallogr. A* **Feb. 1990**, *46 (Pt 2)*, 73–82.

[16] C. R. Kissinger, D. K. Gehlhaar, D. B. Fogel, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 1999**, *55*, 484–491.

[17] N. M. Glykos, M Kokkinidis, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 2000**, *56*, 169–174.

[18] R. J. Read, en, *Acta Crystallogr. D Biol. Crystallogr.* **Oct. 2001**, *57*, 1373–1382.

[19] M. G. Rossmann, D. M. Blow, *Acta Crystallogr.* **Jan. 1962**, *15*, 24–31.

[20] M. Bayes, M. Price, *Philosophical Transactions of the Royal Society of London* **Jan. 1763**, *53*, 370–418.

[21] L. C. Storoni, A. J. McCoy, R. J. Read, en, *Acta Crystallogr. D Biol. Crystallogr.* **Mar. 2004**, *60*, 432–438.

[22] B. C. Wang, en, *Methods Enzymol.* **1985**, *115*, 90–112.

[23] V. Y. Lunin, *Acta Crystallogr. A* **Mar. 1988**, *44*, 144–150.

[24] G. M. Sheldrick, *Zeitschrift für Kristallographie - Crystalline Materials* **Jan. 2002**, *217*, 371.

[25] T. C. Terwilliger, en, *Acta Crystallogr. D Biol. Crystallogr.* **Aug. 2000**, *56*, 965–972.

[26] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2012**, *68*, 352–367.

[27] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2011**, *67*, 355–367.

[28] G. M. Sheldrick, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2010**, *66*, 479–485.

[29] V. S. Lamzin, A Perrakis, K. S. Wilson, *International Tables for Crystallography* **2001**, 720–722.

[30] T. Terwilliger, en, *J. Synchrotron Radiat.* **Jan. 2004**, *11*, 49–52.

[31] K. Cowtan, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept. 2006**, *62*, 1002–1011.

[32] B. Qian, S. Raman, R. Das, P. Bradley, A. J. McCoy, R. J. Read, D. Baker, en, *Nature* **Nov. 2007**, *450*, 259–264.

[33] D. J. Rigden, R. M. Keegan, M. D. Winn, en, *Acta Crystallogr. D Biol. Crystallogr.* **Dec. 2008**, *64*, 1288–1291.

[34] R. Das, D. Baker, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 2009**, *65*, 169–175.

[35] P. E. Leopold, M Montal, J. N. Onuchic, en, *Proc. Natl. Acad. Sci. U. S. A.* **Sept. 1992**, *89*, 8721–8725.

[36] C. B. Anfinsen, en, *Science* **July 1973**, *181*, 223–230.

[37] C. Levinthal, *Mossbauer spectroscopy in biological systems* **1969**, *67*, 22–24.

[38] M. Karplus, en, *Nat. Chem. Biol.* **June 2011**, *7*, 401–404.

[39] Wikipedia, Folding Funnel — Wikipedia, The Free Encyclopedia, [Online; accessed 09-April-2018], **2004**.

[40] J. Lee, P. L. Freddolino, Y. Zhang in *From Protein Structure to Function with Bioinformatics (2nd Ed.) Vol. 69*, (Ed.: D. J. Rigden), Springer Netherlands, Dordrecht, **2017**, pp. 3–35.

[41] J. Skolnick, en, *Curr. Opin. Struct. Biol.* **Apr. 2006**, *16*, 166–171.

[42] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, en, *Methods Enzymol.* **2004**, *383*, 66–93.

[43] D. Xu, Y. Zhang, en, *Proteins* **July 2012**, *80*, 1715–1735.

[44] M. Blaszczyk, M. Jamroz, S. Kmiecik, A. Kolinski, en, *Nucleic Acids Res.* **July 2013**, *41*, W406–11.

[45] T. Kosciolek, D. T. Jones, en, *PLoS One* **Mar. 2014**, *9*, e92197.

[46] S. H. P. de Oliveira, E. C. Law, J. Shi, C. M. Deane, en, *Bioinformatics* **Nov. 2017**, DOI `10.1093/bioinformatics/btx722`.

[47] J. Abbass, J.-C. Nebel, en, *BMC Bioinformatics* **Apr. 2015**, *16*, 136.

[48] Y. Shen, G. Picord, F. Guyon, P. Tuffery, en, *PLoS One* **Nov. 2013**, *8*, e80493.

[49] S. C. Li, D. Bu, X. Gao, J. Xu, M. Li, en, *Bioinformatics* **July 2008**, *24*, i182–9.

[50] I. Kalev, M. Habeck, en, *Bioinformatics* **Nov. 2011**, *27*, 3110–3116.

[51] D. Bhattacharya, B. Adhikari, J. Li, J. Cheng, en, *Bioinformatics* **July 2016**, *32*, 2059–2061.

[52] T. Wang, Y. Yang, Y. Zhou, H. Gong, en, *Bioinformatics* **Mar. 2017**, *33*, 677–684.

[53] S. H. P. de Oliveira, J. Shi, C. M. Deane, en, *PLoS One* **Apr. 2015**, *10*, e0123998.

[54] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, D. Baker, en, *PLoS One* **Aug. 2011**, *6*, e23294.

[55] N Metropolis, S Ulam, en, *J. Am. Stat. Assoc.* **Sept. 1949**, *44*, 335–341.

[56] D Shortle, K. T. Simons, D Baker, en, *Proc. Natl. Acad. Sci. U. S. A.* **Sept. 1998**, *95*, 11158–11162.

[57] Y. Zhang, J. Skolnick, *J. Comput. Chem.* **2004**, *25*, 865–871.

[58] P. Bradley, K. M. S. Misura, D. Baker, en, *Science* **Sept. 2005**, *309*, 1868–1871.

[59] S Ołdziej, C Czaplewski, A Liwo, M Chinchio, M Nanias, J. A. Vila, M Khalili, Y. A. Arnautova, A Jagielska, M Makowski, H. D. Schafroth, R Kaźmierkiewicz, D. R. Ripoll, J Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson, H. A. Scheraga, en, *Proc. Natl. Acad. Sci. U. S. A.* **May 2005**, *102*, 7547–7552.

[60] Y. Zhang, J. Skolnick, en, *Proc. Natl. Acad. Sci. U. S. A.* **May 2004**, *101*, 7594–7599.

[61] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, en, *Nat. Methods* **Jan. 2015**, *12*, 7–8.

[62] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, A. Tramontano, en, *Proteins* **Sept. 2016**, *84 Suppl 1*, 349–369.

[63] C.-H. Tai, H. Bai, T. J. Taylor, B. Lee, en, *Proteins* **Feb. 2014**, *82 Suppl 2*, 57–83.

[64] Z. He, M. Alazmi, J. Zhang, D. Xu, en, *PLoS One* **Sept. 2013**, *8*, e74006.

[65] L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao, N. V. Grishin, en, *Proteins* **Oct. 2011**, *79 Suppl 10*, 59–73.

[66] O. F. Lange, P. Rossi, N. G. Sgourakis, Y. Song, H.-W. Lee, J. M. Aramini, A. Ertekin, R. Xiao, T. B. Acton, G. T. Montelione, D. Baker, en, *Proc. Natl. Acad. Sci. U. S. A.* **July 2012**, *109*, 10873–10878.

[67] S. Raman, O. F. Lange, P. Rossi, M. Tyka, X. Wang, J. Aramini, G. Liu, T. A. Ramelot, A. Eletsky, T. Szyperski, M. A. Kennedy, J. Prestegard, G. T. Montelione, D. Baker, en, *Science* **Feb. 2010**, *327*, 1014–1018.

[68] C. Göbl, T. Madl, B. Simon, M. Sattler, en, *Prog. Nucl. Magn. Reson. Spectrosc.* **July 2014**, *80*, 26–63.

[69] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, C. Sander, en, *PLoS One* **Dec. 2011**, *6*, e28766.

[70] M. Michel, S. Hayat, M. J. Skwark, C. Sander, D. S. Marks, A. Elofsson, en, *Bioinformatics* **Sept. 2014**, *30*, i482–8.

[71] S. Ovchinnikov, L. Kinch, H. Park, Y. Liao, J. Pei, D. E. Kim, H. Kamisetty, N. V. Grishin, D. Baker, en, *Elife* **Sept. 2015**, *4*, e09248.

[72] S. Ovchinnikov, D. E. Kim, R. Y.-R. Wang, Y. Liu, F. DiMaio, D. Baker, en, *Proteins* **Sept. 2016**, *84 Suppl 1*, 67–75.

[73] M. Michel, D. Menéndez Hurtado, K. Uziela, A. Elofsson, *Bioinformatics* **2017**, *33*, i23–i29.

[74] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, D. Baker, en, *Science* **Jan. 2017**, *355*, 294–298.

[75] S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, en, *PLoS Comput. Biol.* **Jan. 2017**, *13*, e1005324.

[76] W. R. Taylor, K Hatrick, en, *Protein Eng.* **Mar. 1994**, *7*, 341–348.

[77] U Göbel, C Sander, R Schneider, A Valencia, en, *Proteins* **Apr. 1994**, *18*, 309–317.

[78] E Neher, en, *Proc. Natl. Acad. Sci. U. S. A.* **Jan. 1994**, *91*, 98–102.

[79] I. N. Shindyalov, N. A. Kolchanov, C Sander, en, *Protein Eng.* **Mar. 1994**, *7*, 349–358.

[80] D. D. Pollock, W. R. Taylor, en, *Protein Eng.* **June 1997**, *10*, 647–657.

[81] A. S. Lapedes, B. Giraud, L. Liu, G. D. Stormo, en in *Statistics in molecular biology and genetics*, Institute of Mathematical Statistics, **1999**, pp. 236–256.

[82] A. Lapedes, B. Giraud, C. Jarzynski, **July 2012**.

[83] F. Simkovic, S. Ovchinnikov, D. Baker, D. J. Rigden, *IUCrJ* **May 2017**, *4*, 291–300.

[84] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, en, *Proc. Natl. Acad. Sci. U. S. A.* **Jan. 2009**, *106*, 67–72.

[85] L. Burger, E. van Nimwegen, en, *PLoS Comput. Biol.* **Jan. 2010**, *6*, e1000633.

[86] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, C. J. Langmead, en, *Proteins* **Apr. 2011**, *79*, 1061–1078.

[87] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, en, *Proceedings of the National Academy of Sciences* **Dec. 2011**, *108*, E1293–E1301.

[88] D. T. Jones, D. W. A. Buchan, D. Cozzetto, M. Pontil, *Bioinformatics* **Jan. 2012**, *28*, 184–190.

[89] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **Jan. 2013**, *87*, 012707.

[90] H. Kamisetty, S. Ovchinnikov, D. Baker, *Proceedings of the National Academy of Sciences* **Sept. 2013**, *110*, 15674–15679.

[91] S. Seemayer, M. Gruber, J. S??ding, en, *Bioinformatics* **Nov. 2014**, *30*, 3128–3130.

[92] T. A. Hopf, D. S. Marks in *From Protein Structure to Function with Bioinformatics*, (Ed.: D. J. Rigden), Springer Netherlands, Dordrecht, **2017**, pp. 37–58.

[93] R. R. Stein, D. S. Marks, C. Sander, en, *PLoS Comput. Biol.* **July 2015**, *11*, e1004182.

[94] T. A. Hopf, S. Morinaga, S. Ihara, K. Touhara, D. S. Marks, R. Benton, en, *Nat. Commun.* **Jan. 2015**, *6*, 6077.

[95] S. D. Dunn, L. M. Wahl, G. B. Gloor, en, *Bioinformatics* **Feb. 2008**, *24*, 333–340.

[96] D. S. Marks, T. A. Hopf, C. Sander, en, *Nat. Biotechnol.* **Nov. 2012**, *30*, 1072–1080.

[97] J. Andreani, J. Söding, en, *Bioinformatics* **June 2015**, *31*, 1729–1737.

[98] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, D. S. Marks, en, *Cell* **June 2012**, *149*, 1607–1621.

[99] M. J. Skwark, D. Raimondi, M. Michel, A. Elofsson, en, *PLoS Comput. Biol.* **Nov. 2014**, *10*, e1003889.

[100] D. T. Jones, T. Singh, T. Kosciolek, S. Tetchner, en, *Bioinformatics* **Apr. 2015**, *31*, 999–1006.

[101] T. Du, L. Liao, C. H. Wu, B. Sun, en, *Methods* **Nov. 2016**, *110*, 97–105.

[102] A. J. González, L Liao, C. H. Wu, *Bioinformatics* **2013**.

[103] G. Shackelford, K. Karplus, en, *Proteins* **2007**, *69 Suppl 8*, 159–164.

[104] J. Cheng, P. Baldi, en, *Bioinformatics* **June 2005**, *21 Suppl 1*, i75–84.

[105] H. Zhang, Q. Huang, Z. Bei, Y. Wei, C. A. Floudas, en, *Proteins: Struct. Funct. Bioinf.* **Mar. 2016**, *84*, 332–348.

[106] Z. Wang, J. Xu, en, *Bioinformatics* **July 2013**, *29*, i266–73.

[107] J. Ma, S. Wang, Z. Wang, J. Xu, en, *Bioinformatics* **Nov. 2015**, *31*, 3506–3513.

[108] B. Adhikari, J. Hou, J. Cheng, en, *Bioinformatics* **Dec. 2017**, DOI `10.1093/bioinformatics/btx781`.

[109] B. He, S. M. Mortuza, Y. Wang, H. B. Shen, Y. Zhang, en, *Bioinformatics* **Mar. 2017**, *33*, 2296–2306.

[110] M. Michel, M. J. Skwark, D. M. Hurtado, M. Ekeberg, A. Elofsson, en, *Bioinformatics* **Sept. 2017**, *33*, 2859–2866.

[111] S. H. P. De Oliveira, J. Shi, C. M. Deane, en, *Bioinformatics* **Feb. 2017**, *33*, 373–381.

[112] Q. Wuyun, W. Zheng, Z. Peng, J. Yang, *Brief. Bioinform.* **Oct. 2016**, bbw106.