Preprint. Final manuscript published in: ***Biosemiotics* 7(2):223-247 (2014)**
doi: 10.1007/s12304-013-9193-5

Biosemiotics manuscript No.
(will be inserted by the editor)

Molecular Codes Through Complex Formation in a Model of the Human Inner Kinetochore

Dennis Görlich · Gabi Escuela · Gerd Gruenert · Peter Dittrich* · Bashar Ibrahim*

Received: date / Accepted: date

This work was supported by the German Research Foundation priority programs InKoMBio (SPP 1395, Grant DI 852/10-1), the European Commission NeuNeu Project (248992) and the Jena School for Microbial Communication (JSMC).

Dennis Görlich
Institute of Biostatistics and Clinical Research, University of Muenster, Muenster, Germany
Tel: +49-251-8353605
E-mail: dennis.goerlich@uni-muenster.de

Gabi Escuela · Gerd Gruenert · Peter Dittrich
Bio Systems Analysis Group, Institute of Computer Science, Jena Center for Bioinformatics and Friedrich Schiller University Jena, Jena, Germany
Tel.:+49-3641-9-46460
Fax: +49-3641-9-46452
E-mail: peter.dittrich@uni-jena.de

Bashar Ibrahim
Bio Systems Analysis Group, Institute of Computer Science, Jena Center for Bioinformatics, Friedrich Schiller University Jena, Jena, Germany, and German Cancer Research Center, DKFZ-ZMBH Alliance, Heidelberg, Germany
Tel.:+49-3641-9-46463
Fax: +49-3641-9-46452
E-mail: bashar.ibrahim@uni-jena.de

Abstract We apply molecular code theory to a rule-based model of the human inner kinetochore and study how complex formation in general can give rise to molecular codes. We analyze 105 reaction networks generated from the rule-based inner kinetochore model in two variants: with and without dissociation of complexes. Interestingly, we found codes only when some but not all complexes are allowed to dissociate. We show that this is due to the fact that in the kinetochor model proteins can only bind at kinetochores by attaching to already attached proteins and cannot form complexes in free solution. Using a generalized linear mixed model we study which centromere protein (CENP) can take which role in a molecular code (sign, meaning, context). By this, associations between CENPs (CenpA, CenpQ, CenpU and CenpI) and code roles are found. We observed that CenpA is a major risk factor (increases probability for code role) while CenpQ is a major protection factor (decreases probability for code role). Finally we show, using an abstract model of copolymer formation, that molecular codes can also be realized solely by the formation of stable complexes, which do not dissociate. For example, with particular dimers as context a molecular code mapping from two different monomers to two particular trimers can be realized just by non-selective complex formation. We conclude that the formation of protein complexes can be utilized by the cell to implement molecular codes. Living cells thus facilitate a subsystem allowing for an enormous flexibility in the realization of mappings, which can be used for specific regulatory processes, e.g. via the context of a mapping.

Keywords Molecular codes · Modeling approaches · Inner kinetochore · Rule-based modeling · Generalized linear mixed models · S-phase

Mathematics Subject Classification (2000) 92-08 · 92C37 · 92C42 · 62P10

Introduction

Cells implement a large variety of information processing subsystems, which play a central role in fundamental processes like evolution, communication, regulation, and adaptive behavior (Holcombe and Patton 1998). To understand such cellular information processing systems formal methods like graph theory (Farkas et al 2001), dynamical systems theory (Klipp et al 2010), and information theory (Waltermann and Klipp 2011; Harmer 2010) are often applied. However, these methods do not consider semantic nor pragmatic aspects of biological information (Shannon 1948; Klipp et al 2010). A series of contributions discuss how semiotics, the science of signs, could lead to a deeper and more unified understanding of biological information (Sebeok 2001; Barbieri 2008b; Favareau 2010). From these studies it is becoming clear that "organic" codes play a central role in many processes and are a feasible instrument to get a better understanding of semantic aspects of biological information (Barbieri 2008a).

In a previous work we have suggested a formal concept to detect molecular codes in reaction networks (Görlich and Dittrich 2013). Roughly, given a reaction network, a *molecular code* is a mapping from a subset of molecular species to a subset of molecular species (called signs and meanings) provided that the network is able to realize the mapping (using a particular subset of molecular species called context) and that the network is able to realize a *different* mapping from the same signs to the same meanings, using a different subset of molecular species called alternative context. The latter property makes the mapping "contingent" or "arbitrary", i.e. it can be different by changing the context. We have chosen the reaction network as a formal and experimentally verifiable description of a physical system of interest. Thus, whether a mapping between molecules is contingent or not can – at least in principle – be experimentally determined.

Applying a software tool that is able to find molecular codes in reaction networks, we looked for molecular codes in non-biological networks, like various combustion chemistries and a Martian atmospheric photochemistry, and in abstract biological networks describing translation, signaling, and gene regulation. Interestingly, hardly no molecular codes were found in the non-biological networks, while a large number of molecular codes where found in the biological networks (Görlich and Dittrich 2013). This suggests that life has acquired access to a chemistry with a relative high potential to realize molecular codes. With our formal molecular code concept we can now even formulate an experimentally testable and quantifiable hypothesis that the ability to realize molecular codes has increased over life's evolution.

In order to get hold on this idea it is necessary to study what kind of mechanisms can give rise to molecular codes and to investigate more concrete biological systems in detail. Therefore, in this work we will look at complex formation as a potential mechanism for code generation and investigate a concrete rule-based reaction network model of the kinetochore by Tschernyschkow et al (2013). A kinetochore is a multi protein complex that forms at the centromere of each sister chromatid (Gascoigne and Cheeseman 2010). As opposed to our previous study of more abstract biological networks (Görlich and Dittrich 2013), this is the first analysis of a concrete bio-molecular system.

In the reminder of the introduction we will briefly review the biological background of the kinetochore model and our molecular code concept. Then we describe our methods for constructing and analyzing the kinetochor model. In the result section we present the analysis of 210 networks sampled from the rule-based model in two variants: 105 networks with dissociation and 105 networks without allowing complexes to dissociate. We show under which conditions molecular codes appear and which proteins are involved in which role. In particular, in the kinetochor model we found codes only when the reaction network contains some dissociation reactions. We observed that CenpA is a major risk factor (increases probability for code role) while CenpQ is a major protection factor (decreases probability for code role).

In the discussion we explain the observed codes by an abstract reaction network model implementing the essential mechanism. Finally we show by an

abstract model of polymerization that molecular codes can also be realized solely by unconstrained complex formation without dissociation.

Description of the inner kinetochore rule-based model

Faithful chromosome segregation is mediated by a multi protein complex which assembles solely at the centromere of each sister chromatid, called the kinetochore. Its malfunctioning results in aneuploidy and can lead to development of cancer (Cimini and Degrassi 2005; Suijkerbuijk and Kops 2008; Holland and Cleveland 2009; Li et al 2009). A kinetochore contains over 100 proteins and complexes. These proteins can be classified into two functional parts: the inner and the outer kinetochore. The outer kinetochore is less stable and forms in early mitosis (Maiato et al 2004; Cheeseman and Desai 2008) while the inner kinetochore is more stable and present during the entire cell cycle (Dalal and Bui 2010; Black and Cleveland 2011; Perpelescu and Fukagawa 2011). Inner kinetochore proteins and complexes are conserved throughout evolution including a centromeric CenpA and 16 CCAN proteins (CenpC, CenpH, CenpI, CenpK to CenpU, CenpW, CenpX) (Okada et al 2006).

Studying the 3D structure of the kinetochore is challenging, both experimentally and theoretically, because of the combinatorial explosion of the number of intermediate complexes (Tschernyschkow et al 2013). Explicit representations of all intermediate molecular species, e.g. as system of differential equations, Boolean networks or Bayesian networks, cannot account for this combinatorial explosion. The reason is that these modeling approaches use a restricted state space with fixed dimensionality. Recently, the inner kinetochore has been modeled (Tschernyschkow et al 2013), based on an implicit representation, which combines a rule-based description language and spatial aspects (Gruenert et al 2010). The model is developed based on intra-cellular proximity between inner kinetochore proteins by measuring the Förster resonance energy transfer (FRET) in addition to literature data (Tschernyschkow et al 2013). This approach suites to cope with combinatorial complexity and includes molecular geometric information in contrast to classical modeling approaches applied so far to cell-cycle mechanisms (e.g. (Doncic et al 2005; Lohel et al 2009; Ibrahim et al 2008b, 2009, 2008a, 2007; Ibrahim 2008; Rohn et al 2008; Caydasi et al 2012)).

In this study, we applied the novel molecular code theory (Görlich and Dittrich 2013) to the human inner kinetochore model recently published by Tschernyschkow et al (2013). We considered two model variants, with and without dissociation reactions, respectively. We determine and analyze all possible codes as well as relate these codes to biological well characterized functions.

The molecular code framework

In Görlich and Dittrich (2013) we have introduced a formal definition of molecular code with respect to a reaction network as a contingent mapping on

molecular species. A reaction network is given by a set of molecular species, e.g. $\{A, B, C, D, E, F, G, H\}$, and a set of reaction rules, e.g. $\{A+E \to C, A+F \to D, B+G \to C, B+H \to D\}$.

A reaction rule means that if in a reaction vessel the molecules of the left hand side are present, eventually the right hand side molecules can appear, i.e. the reaction takes place. Using this reaction network we can realize a mathematical mapping from one set of molecular species to another one. Consider, for example, the mapping $f : \{E, F\} \to \{C, D\}$ with $f(E) = C$ and $f(F) = D$. We can realize this mapping by a reaction vessel containing molecules of type A (called context). To compute $f(E)$ we add E to the vessel, wait for some while, and check whether C or D appears. Notice that the network cannot realize a different mapping on the same domain and codomain, i.e. it cannot realize the mapping $f' : \{E, F\} \to \{C, D\}$ with $f'(E) = D$, $f'(F) = C$. Thus the mapping f is not contingent and thus not a molecular code.

Now consider the mapping $g : \{A, B\} \to \{C, D\}$ with $g(A) = C$ and $g(B) = D$. This mapping is a molecular code because it can be realized by the network using the context $\{E, H\}$ and because there is a different mapping $g' : \{A, B\} \to \{C, D\}$ with $g'(A) = D$ and $g'(B) = C$ on the same domain and codomain that can be realized by the network using the alternative context $\{F, G\}$

Instead of analyzing arbitrary large codes, i.e. codes with arbitrary large domains and codomains, we restrict our analysis without loss of generality to binary molecular codes. A binary molecular code (BMC) is a contingent mapping from two signs to two meanings, like our example above. It is possible to merge several BMCs to obtain larger codes. Thus, for a general analysis, to answer the question if codes could be realized in a system, the identification of BMCs is sufficient.

The genetic code, for example, is a molecular code that maps codons (domain) to each amino acid (codomain). The mapping is realized by the appropriate tRNAs (contexts) (Görlich and Dittrich 2013). The biochemical basis is given by the modularity of the tRNA, or to be more specific the biochemical process that loads an amino acid on the tRNA.

Material and Methods

We applied a set of analysis techniques described in this section and illustrated in Fig. 1.

Rule-based model of the inner kinetochore

As basic model of the inner kinetochore we used the rule-based model (Fig. 2) proposed by Tschernyschkow et al (2013). The model is specified in the BioNetGen Language (BNGL) (Faeder et al 2009). In BNGL each Cenp and

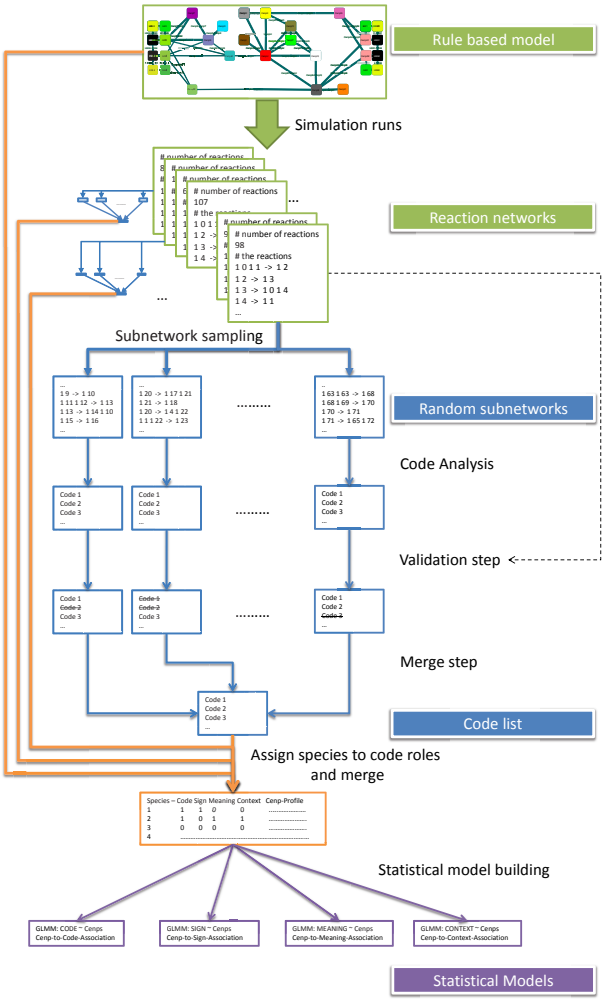

Fig. 1 Steps from the rule-based model to the statistical models. The rule-based model is simulated repeatedly and the instantiated reaction networks are observed. These are analyzed for codes using a random subnetwork sampling approach. Identified codes have to be validated in the full reaction network. Validated codes are pooled and species used in codes are determined. Over all reaction networks species-to-code data is merged and supplemented by the Cenp-profile of each species. Generalized linear mixed models (GLMM) are used to statistically model the Cenp-to-code-role associations.

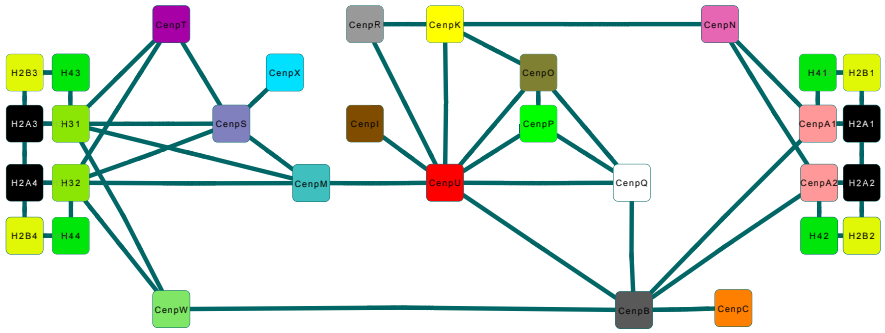

Fig. 2 Schematic network of human inner kinetochore model (redrawn and adapted from (Tschernyschkow et al 2013)). The vertices refers to the molecules, while the edges refers to the binding data (FRET proximities). The two nucleosomes (left and right) are the anchor points for the model calculations. The CenpA containing nucleosome contains CenpA1, CenpA2, H41, H42, H2A1, H2A2, H2B1, H2B2. The H3 containing nucleosome contains H31, H32, H43, H44, H2A3, H2A4, H2B3, H2B4.

Histone is specified by its potential binding sites, and thus, implicitly, a reaction network of the model is defined. For the code based analysis we need to generate the reaction network, i.e. the explicit representation of the model.

Observing a reaction network

There are two ways to obtain a network model from the rule-based description of the inner kinetochore. In theory it is possible to analytically derive all possible reaction rules from the binding rules describing the system. This approach showed to be computational not feasible for the inner kinetochore model due to the combinatorial complexity on the number of potential protein complexes and on the number of reaction among these complexes. We, therefore, chose to generate networks by "observing" simulation runs of the system. On the one hand, this allows to concentrate our analysis on the reactions important for the function of the kinetochore, because these emerge in most of the simulations, on the other hand we might "miss" code forming mechanisms, due to incomplete coverage. Basically, incorporating all reactions includes many reaction paths that might not be directly relevant and unnecessarily complicate the code identification. We applied the SRSim software developed by Gruenert et al (2010) and modified it accordingly to observe the reactions in each step of the simulation. One simulation runs for 3×10^6 time steps, which proved to be enough for the model to form bridges between two nucleosomes (Tschernyschkow et al 2013). Because simulations are not deterministic the resulting network models differs from each other in number of molecular species and number of reactions. We, therefore, repeated the simulations. In total we ran 105 simulations with dissociation and 105 simulations without dissociation reactions.

Code analysis

A network model can be analyzed algorithmically for BMCs (Görlich and Dittrich 2013). The algorithm checks all pairs of molecular species against all (other) pairs of molecular species, if they constitute a molecular code by identifying the molecular contexts and checking the code conditions as defined in (Görlich and Dittrich 2013). This simple, brute force, algorithm performs bad on large networks because of the fast growing number of pairs, but ever worse on the faster growing number of paths through the network. We, therefore, developed a subnetwork sampling algorithm that uses random subnetworks and searches for codes in these. If a BMC is found in a subnetwork it has to be validated in the complete network. The validation step is computationally not expensive and the whole algorithm allows to analyze larger networks. Parameters that can be chosen for the algorithm are: K - the number of shortest paths considered for the code identification; S - the approximated size of the sampled subnetworks; rep - the number of subnetworks sampled . For this analysis we empirically determined K=1, S=50 and rep = 1000 as suited parameters. Larger subnetworks lead to longer running times, as do larger values of K. The number of repeats should yield a reasonable coverage of the original network. The general procedure is shown in Algorithm 1.

Algorithm 1 Subnetwork sampling algorithm

1: **Choose** a random molecular species
2: **Expand** the subnetwork N_{sub} around this species until threshold S is reached
3: **Analyze** N_{sub} using the code finding algorithm (Görlich and Dittrich 2013) with parameter K
4: **Validate** all found codes in N
5: **Repeat** 1 - 4 until the maximal number of repeats rep is reached
6: **Report** all validated codes

In each repetition in Step 1 a random molecular species is drawn. Around this species, following incoming and outgoing reactions, the subnetwork is expanded (Step 2), using a closure operator (cp. (Görlich and Dittrich 2013)) until the size of the subnetwork exceeds the predefined threshold S. Due to the closure operator the number of molecular species will usually be larger than S because in one expansion step more than one species is included into the subnetwork. Step 3 in the code analysis workflow uses the code identifying algorithm proposed in (Görlich and Dittrich 2013). Basically, all possible combinations of molecular species are tested for the code property for binary molecular codes (BMC). The code property guarantees that uniqueness of two contingent mappings is not violated by the network. In Step 4 identified codes have to be validated in the complete network model. This step is necessary because reactions that are not part of the subnetwork, can (and will) destroy the code property (mainly the uniqueness of the mapping). All validated BMCs can be reported (Step 6).

The algorithm does not guarantee to identify all codes, but converges towards to the real number if the number of samples is large.

Statistical analysis

To test for associations between CENPs and code roles we are pooling the data generated by analyzing the independent simulation runs. To account for the repeated occurrence of the molecular species we apply a generalized linear mixed model (GLMM) which can be used for repeated measurement data or, as in our case, clustered data (multiple observations of the same molecular species). To account for the binary response variables we use a binomial distribution and a logit link. The resulting (exponentiated) model coefficients, therefore, can be interpreted as odds ratios. The model coefficients can also be visualized as forest plots showing the odds ratios and the 95%-confidence interval (CI) (see Appendix). CENPs, whose CI is not containing 1, are considered to have a significant association with the respective code role.

To assess model qualities we predicted the mean response for each case and used it in a ROC-Analysis as predictor for the actual target variable (CODE, SIGN, MEANING, CONTEXT). Table 1 displays the obtained area under the curve (AUC) values. All the observed AUC values are larger than 0.8, which indicates a good predictive behavior of the fitted models. Significance values (p-values) show that the classification performance is significantly better than a random classifier. Model fits were generated using SAS/STAT software, Version 9.2 of the SAS System for Windows. Copyright © 2008 SAS Institute Inc. ROC-Analysis has been performed with SPSS (IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0., Armonk, NY). Forest plots have been generated with R 2.15.1 using package `rmeta` (version 2.16).

Table 1 Model performance measured by a ROC analysis of the predicted mean values against the true value.

				95%-CI	
Model	AUC	StdError	Significance	Lower	Upper
CODE	0.848	0.004	<0.001	0.840	0.855
SIGN	0.819	0.005	<0.001	0.810	0.829
MEANING	0.865	0.004	<0.001	0.857	0.874
CONTEXT	0.931	0.003	<0.001	0.925	0.936

AUC - Area under the ROC-curve; ROC - receiver operator characteristic; CI - confidence interval.

Results

We analyzed 105 networks allowing for dissociation of the protein complexes and 105 networks without dissociation. Generation of the network models is

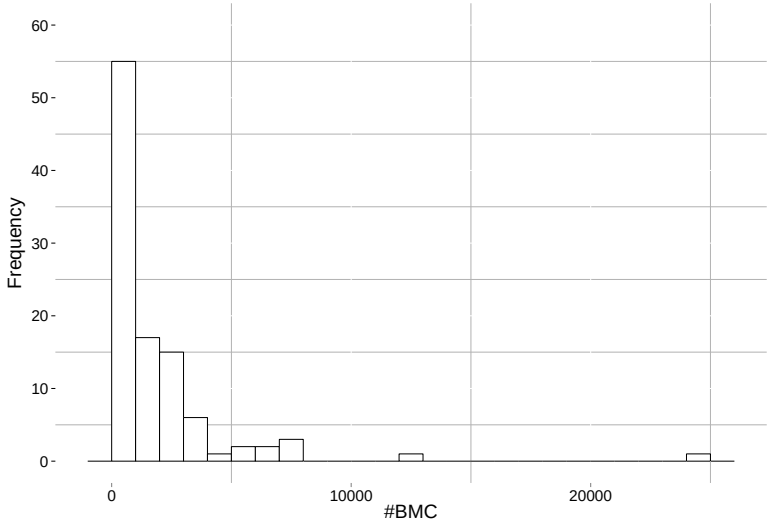

Fig. 3 Distribution of the number of BMCs in the 105 analyzed networks with dissociation.

described in more detail in the Material and Methods section. The application of the heuristic algorithm to these networks shows that without dissociation no codes can be realized in the inner kinetochore formation. With dissociation, the number of BMCs ranges from 0 to 24,315 BMCs with a median of 748 (Fig. 3).

Association of CENPs to code roles

A code consists of a set of molecular species that can be denoted as signs, a set of meanings and a set of molecular contexts that realizes the mapping. Here we analyze the molecular species present in the simulated networks for their association with these three roles. Each species has a specific Cenp profile, i.e. the combination of different CENPs. We are using this Cenp profile as covariates to fit a statistical model explaining the association between the single CENPs and the role a molecular species takes. To include the data of all 105 networks and to account for possible correlations (e.g. when the same species is present several times) we chose a general linear mixed model (GLMM). The dataset consists of 9257 cases representing 5283 different molecular species that are repeatedly observed between 1 and 84 times, i.e. different networks. We consider 16 input variables, i.e. the count of Cenp proteins, and 4 target variables. The target variables (CODE, SIGN, MEANING, CONTEXT) are binary and indicate if the respective species was part of (1) a code (in any role), (2) a sign, (3) a meaning, or (4) a context. For the analysis the histones have been excluded, since they only are important as a scaffold, but not of interest for the code analysis. We estimate four models, i.e. one model for

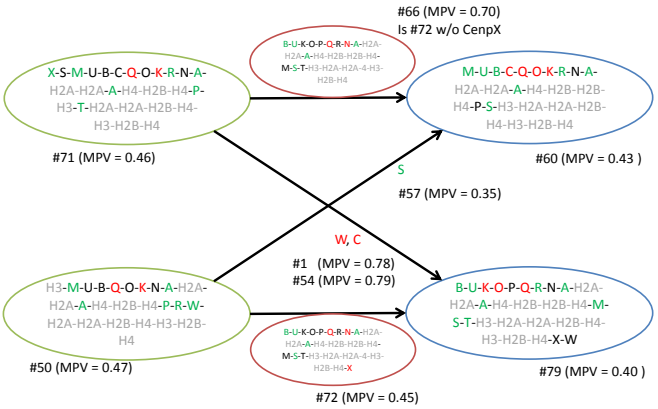

Fig. 4 An exemplary code. Two signs on the left hand side, two meanings on the right hand side. The two possible mappings are overlaid and annotated with the respective molecular context. Each molecular species is identified by its combination of CENPs and histones. The mean predicted value (MPV) gives the GLMM model prediction for the protein complex to have the respective code role. MPV values, here, range between 0 and 1. Colors refer to the result of the Cenp-to-role association (see next sections). Green - risk factor (increases probability for that role), red - preventive factor (decreases probability for that role), grey - histones (have not been included in the models), black - no significant association could be found. MPV - mean predicted value.

each target variable, using a binomial distribution, accounting for the binary response variables, and a logic link. In each model we calculate the odds ratios (OR) as measure for a molecular species' association with the respective code roles. Each OR is accompanied by a statistical test indicating if it differs from 1 (no effect) significantly. We consider an OR as significant when $p \leq 0.05$. A significant OR > 1 indicates a risk factor, i.e. an increase in the respective co-variable increases the probability to be, e.g. part of a code. An OR < 1 indicates a "protective" factor, i.e. an increase in this variable decreases the probability to be, e.g. part of a code. To get a better overview of the results we use forest plots to visualize odd ratios and their 95%-confidence intervals. Model fits have been calculated with SAS9.2 (SAS Software 9.2, TS1M0, SAS Institute Inc., Cary, NC, USA).

CENPs associated with codes

The variable CODE is a combined endpoint from all code roles, i.e. a species being present in at least one role is considered to be part of a code. The GLMM of the dataset with respect to the usage of species in codes in general shows that all CENPs are associated with the endpoint CODE. CenpC, -K, -N, -O , -Q, -W, and -X are protective factors that reduce a species chance to be part of a code by a factor 0.76 (s.d. 0.04) on average. CenpA, -B, -I, -M, -P, -R, -S, -T, and -U are "risk" factors that increase the chance of a molecular species

to be part of a code by a factor 1.48 on average (s.d. 0.25) (see Table 2, Table 3 first column and Fig. 9).

Table 2 Odd ratio estimates of the generalized linear mixed model for the response CODE.

						95%-CI	
	Protein	Estimate	StdErr	Significance	OR	Lower	Upper
1	**CenpA**	0.17	0.05	**0.0003**	**1.18**	**1.08**	**1.29**
2	**CenpB**	0.38	0.10	**0.0001**	**1.47**	**1.21**	**1.79**
3	**CenpC**	-0.27	0.07	**0.0001**	**0.77**	**0.67**	**0.88**
4	**CenpI**	0.36	0.07	**<.0001**	**1.44**	**1.26**	**1.64**
5	CenpK	-0.24	0.10	**0.0131**	**0.78**	**0.64**	**0.95**
6	**CenpM**	0.44	0.10	**<.0001**	**1.55**	**1.27**	**1.91**
7	**CenpN**	-0.24	0.09	**0.0093**	**0.78**	**0.65**	**0.94**
8	**CenpO**	-0.32	0.09	**0.0007**	**0.73**	**0.61**	**0.88**
9	**CenpP**	0.29	0.08	**0.0005**	**1.34**	**1.14**	**1.57**
10	**CenpQ**	-0.37	0.10	**0.0001**	**0.69**	**0.57**	**0.84**
11	**CenpR**	0.51	0.08	**<.0001**	**1.66**	**1.43**	**1.93**
12	**CenpS**	0.68	0.11	**<.0001**	**1.97**	**1.59**	**2.43**
13	**CenpT**	0.29	0.07	**<.0001**	**1.34**	**1.17**	**1.53**
14	CenpU	0.22	0.11	**0.0439**	**1.25**	**1.01**	**1.54**
15	**CenpW**	-0.22	0.07	**0.0014**	**0.80**	**0.70**	**0.92**
16	**CenpX**	-0.17	0.07	**0.0114**	**0.84**	**0.74**	**0.96**

CENPs associated with signs

For the target variable SIGNS the estimated model (Table 4) is more differentiated than for the combined target CODE. We can now observe that CenpA, -I, -M, -P, -R, -T, -W and -X are risk factors increasing a protein complex' probability to be a sign on average by a factor of $1.38(s.d0.23)$, while CenpK and CenpQ can be considered protective factors, reducing a protein complex probability to be a sign by $0.65(CI : 0.54 - 0.78)$ and $0.70(CI : 0.59 - 0.85)$, respectively. Among the risk factors CenpM and CenpI are the strongest factors increasing the "risk" to be a sign by 75% $(CI : 1.47 - 2.15)$ and 70% $(CI : 1.50 - 1.92)$, respectively (see Table 4, Table 3 second column and Fig. 10).

CENPs associated with meanings

Similarly to the analysis of the association of CENPs to endpoint SIGNS, the association to the endpoint MEANING (Table 5) is also more differentiated than for CODE. For MEANING CenpA, B, I, M, R, S, T and U are risk factors (average OR = $1.42(s.d.0.28)$), while CenpC, K, O, and Q are protective factors (average OR = $0.81(s.d.0.04)$). The strongest risk factor is CenpS (OR=2.09,$CI : 1.64 - 2.62$). The strongest protective factor is CenpO (OR=0.77,$CI : 0.64 - 0.93$) (see Table 5, Table 3 third column and Fig. 11).

CENPs associated with contexts

The analysis of the model for the code role CONTEXT (Table 6) reveals CenpA, -B, -S, and -U as risk factors (mean OR = $1.95(s.d = 0.74)$) and CenpC, -I, -N, -Q, -W, and -X as protective factor with a mean OR of $0.52(s.d. = 0.14)$. The strongest risk factor is CenpS with an increase in risk by almost 300% (OR = 2.96, $CI : 2.20 - 3.99$). The strongest protective factor is CenpX reducing the risk by a factor of $0.30(CI : 0.26 - 0.36)$ (see Table 6, Table 3 fourth column and Fig. 12) .

Summary

Table 3 summarizes the observed significant CENPs. CENPs that function as "risk factor" are marked by a green upward arrow, while CENPs that function as "protection factor" are marked by a red downward arrow. For non-significant CENPs no tendency of the OR can be deduced, and thus no arrow is shown. All selected CENPs, except CenpX and CenpS, are consistent, i.e. if they are risk factor for codes, they are also risk factor for a particular code role, or if they are protection factors for codes, they are also protective factor for the other code roles. CenpA, is the major risk factors (OR value > 1) for codes, while CenpQ is the major protection factors (OR value < 1) where their effects can be seen in each of code, sign, meaning and context. CenpB, -I, -M, -R, -S, and -T can be considered as a risk factors (OR value > 1) for at least three of the four code roles (Code, Sign, Meaning, or Context). Similarly, CenpC can be considered as protective factor for the same reason. Only CenpI and CenpX switch their effect between risk and protection.

Relating CENPs to function

The establishment of kinetochores is dependent on the presence of the centromere-specific nucleosome that contains the H3 variant, CenpA (Cse4 in budding yeast, Cnp1 in fission yeast, and CID/CenH3 in fruit flies) (Obuse et al 2004; Howman et al 2000; Stoler et al 1995; Sullivan et al 1994; Topp et al 2004; Bergmann et al 2011). Additionally, bridges between nucleosomes required CenpA to be formed (Tschernyschkow et al 2013). During S-phase, but not M-phase, CenpA nucleosome consider to be "the sole epigenetic mark of centromere" (for review see (Quenet and Dalal 2012; Perpelescu and Fukagawa 2011)). Our code analysis reveals that CenpA is essential as it is a "risk" factor, in other words, CenpA-containing protein complexes are likely to be part of codes.

Human CenpQ binds directly microtubules *in vitro* (Amaro et al 2010). Thus, it has a central role in mitosis and not in S-phase. Our analysis of the S-phase inner kinetochore model shows that CenpQ is a "protective" factor, which means that it is likely not a part of a code. This is in the same concert with the known data.

Table 3 Overview of the effect and significance of the Cenp-Proteins in the respective models. For OR: green arrows (OR $\geq$ 1.2), yellow up (1 $<$ OR $\leq$ 1.2) , yellow down (0.8 $\leq$ OR$<$ 1), red down (OR $<$ 0.8). For p-value: green circle (p $\leq$ 0.05), yellow circle (0.05 $<$ p $\leq$ 0.1, does not occur), red (p $>$ 0.1). p-values displayed as 0.0000 indicate a value smaller that 0.0001.

	CODE		SIGN		MEANING		CONTEXT	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CenpA	1.18	0.0003	1.15	0.0011	1.14	0.0054	1.31	0.0000
CenpB	1.47	0.0001	0.93	0.4660	1.39	0.0017	2.05	0.0000
CenpC	0.77	0.0001	0.94	0.3285	0.87	0.0485	0.56	0.0000
CenpI	1.44	0.0000	1.70	0.0000	1.41	0.0000	0.60	0.0000
CenpK	0.78	0.0131	0.65	0.0000	0.79	0.0230	1.02	0.8998
CenpM	1.55	0.0000	1.75	0.0000	1.38	0.0081	1.23	0.1818
CenpN	0.78	0.0093	1.07	0.4381	1.03	0.7465	0.57	0.0000
CenpO	0.73	0.0007	0.88	0.1447	0.77	0.0064	0.93	0.5628
CenpP	1.34	0.0005	1.43	0.0000	1.10	0.2763	1.08	0.4603
CenpQ	0.69	0.0001	0.71	0.0002	0.81	0.0373	0.70	0.0041
CenpR	1.66	0.0000	1.27	0.0014	1.24	0.0071	1.17	0.1346
CenpS	1.97	0.0000	1.12	0.2766	2.08	0.0000	2.96	0.0000
CenpT	1.34	0.0000	1.26	0.0006	1.41	0.0000	1.04	0.6868
CenpU	1.25	0.0439	1.17	0.1476	1.38	0.0088	1.50	0.0079
CenpW	0.80	0.0014	1.30	0.0000	1.07	0.3094	0.40	0.0000
CenpX	0.84	0.0114	1.17	0.0108	1.08	0.2480	0.30	0.0000

CenpU (also known as MLF1IP or Cenp-50) is a component of CCAN due to its co-localization with CenpA throughout the cell cycle in human cells (Hanissian et al 2004; Cheeseman and Desai 2008; Okada et al 2006). CenpU is required for stable kinetochore-microtubule attachment (Hua et al 2011). Its depletion can cause a mitotic defect in chromosome attachment and chromosome alignment but not affecting the spindle assembly checkpoint (Foltz et al 2006). The only known function for CenpU for the inner kinetochore is the link with CenpA and stabilize it (Hua et al 2011). Our code analysis is supporting these data (Hua et al 2011). CenpU is a risk factor for both, meaning and context. This can be seen as microtubules binding preventing some coding and activates others like CenpU.

CenpI has important roles for the inner and also for the outer kinetochore like spindle assembly checkpoint activity via Ndc80 and Mad1/2 (Liu et al 2003). Our code theory results in switching behavior for CenpI (either as risk or protective factor). We relate this behavior to the limitation of the inner kinetochore model (Tschernyschkow et al 2013) where solely single bound between CenpI and CenpU has been considered. Additionally, a counter partner of CenpI called CenpH is also not in the model (Tschernyschkow et al 2013).

Discussion and Conclusions

Our analysis of the kinetochore model showed that non of the networks without dissociation could realize BMCs. While molecular codes were found when some dissociation reactions were contained in the sampled reaction networks. This is insofar interesting, because complex formation without dissociation can in principle also lead to molecular codes, as we will show bellow (cf. Example 2). Further note that if all complexes could fully dissociate, no molecular codes can be realized by the network, because of lacking closed sets. So, in our kinetochor model an intermediate level of dissociation is required for molecular codes, i.e. some complexes must be stable while some have to be unstable. In the following we explain the underlying mechanisms by an abstract kinetochore and a polymerization model. The important difference is that in the polymerization network complexes can form arbitrarily (i.e. also in free solution) while in the kinetochore model bonds can only form when a binding partner is already part of a nucleosome containing complex.

Example 1: An abstract inner kinetochore model

Let us consider an abstract model in which a bridge between two nucleosomes x and y is formed by three proteins a, b, and c (Fig. 5A). A bond can only be formed if one binding partner is part of a nucleosome containing complex, which is a specific property of the kinetochor model (for justification see above and (Tschernyschkow et al 2013)). The resulting reaction network consists of 12 molecular species and the following reaction rules (not including dissociation reactions):

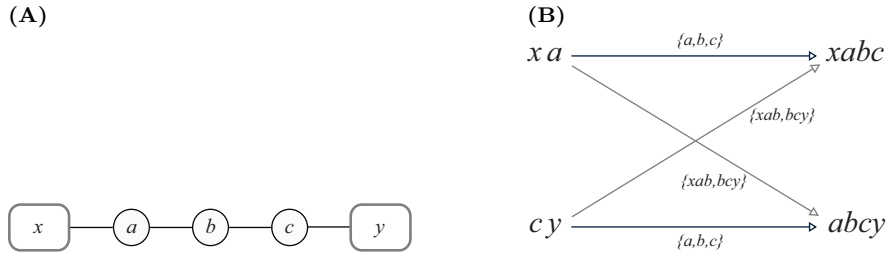

Fig. 5 Simple model of protein complex formation abstracting inner kinetochore formation. In panel (A), x and y represents nucleosomes that are assumed to be fixed. a,b,c are proteins that can only build complexes and form a bridge between the nucleosomes. Additionally, we restrict the binding order, according to the biological binding order observed in the kinetochore, a and c needs to be bound to x or y, respectively, before b can be bound to any subcomplex. Panel (B) shows a code pair found for the variant where two dissociation reactions are allowed, namely, $xa \to x + a$ and $yc \to y + c$. This code pair has $\{xa, cy\}$ as signs and $\{xabc, abcy\}$ as meanings. Note that the arrows, annotated by the respective molecular context, do not denote reactions but the two mappings of the code pair.

$x + a \to xa$	$c + y \to cy$
$xa + b \to xab$	$cy + b \to bcy$
$xab + c \to xabc$	$bcy + a \to abcy$
$xabc + y \to xabcy$	$bacy + x \to xabcy$

We tested for codes in four different variants of the model: without any dissociation, with partial dissociations (single reaction or two reactions), and with full dissociation reactions. We did not find any code except for the variant in which two reactions dissociate. We took as an example the dissociation reaction $xa \to x + a$ and $yc \to y + c$ and we found a code where $\{xa, cy\}$ are the signs and $\{xabc, abcy\}$ are the meanings with context $\{a, b, c\}$ and an alternative context $\{bcy, xab\}$ (Fig. 5B).

Taking together, the inner kinetochore model assumptions for specific ordering in complex formation without dissociation does not allow for molecular codes, while additional dissociation can reintroduce this property. Then, intermediate complexes can function as a kind of carrier for another necessary protein.

Example 2: Polymerization without dissociation

In order to show that molecular codes can also be realized without dissociation, we consider a simple model of the formation of stable copolymers consisting of two different monomers p and q. The reaction network is given by:

$p + p \to pp$	$pq + q \to pqq$
$p + q \to pq$	$p + qq \to pqq$
$q + p \to qp$	$qp + p \to qpp$
$q + q \to qq$	$q + pp \to qpp$
$pp + p \to ppp$	$qp + q \to qpq$
$pp + q \to ppq$	$q + pq \to qpq$
$p + pq \to ppq$	$qq + p \to qqp$
$pq + p \to pqp$	$q + qp \to qqp$
$p + qp \to pqp$	$qq + q \to qqq$

This network (Fig. 6) can realize four code pairs. In these codes the signs and meanings of each code, respectively, are:
Code 0: $\{p, q\} \to \{ppq, pqq\}$ with contexts $\{pq\}, \{pp, qq\}$
Code 1: $\{p, q\} \to \{ppq, qqp\}$ with contexts $\{pp, qq\}, \{pq, qp\}$
Code 2: $\{p, q\} \to \{pqq, qpp\}$ with contexts $\{pq, qp\}, \{pp, qq\}$
Code 3: $\{p, q\} \to \{qpp, qqp\}$ with contexts $\{qp\}, \{pp, qq\}$
Analysis of the nesting structure (cp. (Görlich 2013)) shows that the four codes are not independent from each other, but a more complex interdependence

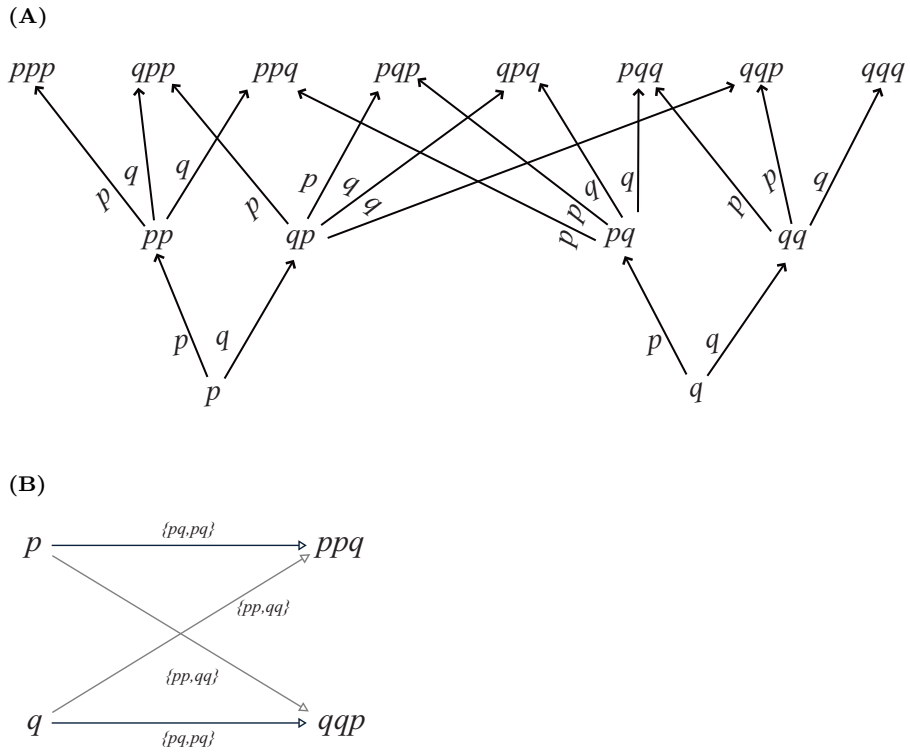

Fig. 6 Graphical representation of the polymerization example. In panel (A) two molecular species, p and q, can polymerize according to the given reaction rules (see text). We assume only two steps in the polymerization, i.e. p and q can form polymers of length 2 and in a second step all polymers of length 3 can form. Panel (B) shows an exemplary code found for the polymerization example

exists, suggesting that the codes share common molecular mechanisms (Fig. 7).

From these two examples, we conclude that reversibility of complex formation reactions is a critical property from molecular codes theory point of view. Thus, it would be interesting to conduct future studies addressing these effects.

Towards the pragmatic level - code usage

Code roles can be interpreted in different ways. First, the assignment of sign and meanings, marks the flow of information in the system. Due to the directed structure of the reaction network model we know that the effects of incoming external[1] signals can only have an effect on downstream targets, e.g. meanings. Such control of the execution of the code can either be achieved via the signs,

[1] External in the sense of the code, i.e. not the signs, or contexts itself.

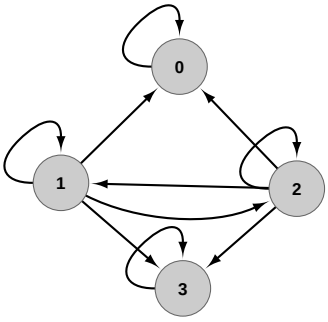

Fig. 7 Code nesting in the polymerization example without dissociation. Arrows point towards the nested code. Codes 1 and 2 are 'mutually' nested, as they share common molecular contexts. A code is always nested in itself (reflexivity).

or the context. A regulation of the context, for example, is consistent with a regulation of the mapping itself, i.e. dynamic changing of the mapping. Presence of signs, on the other hand, is necessary for code execution, and does not have any further regulatory effect.

The code analysis at the actual point of development is useful to identify codes from the structural perspective, addressing the semantic level of cellular information processing. From a dynamic perspective it is necessary that the two contexts are not present at the same time to ensure uniqueness of the mapping. For example, in the genetic code the ribosome guarantees that only one tRNA per time has access to the messenger RNA, while only one codon is presented at once. This represents a type of spatial separation. Code utility can also be achieved by a temporal separation, i.e. differences in the dynamics of the respective molecular species. The cell can realize a temporal separation of contexts, by regulating or controlling the "production" of the molecular species from the domain and the "selected" context in parallel to ensure the "execution" of the mapping.

Code analysis in the context of biological modeling

For the analysis of biological systems a huge variety of techniques is available. Beside purely structural analyzes, like path analysis, the incorporation of dynamic information will be considered for future research. An overview how code theory method relates to classical modeling and future algebraic method is presented in Fig. 8. Analysis techniques are often coupled with the modeling representation of the system in question. In general, we can distinguish between explicit and implicit representations (Fig. 8). While the former models all known reactions and interactions between system components explicitly,

e.g. as reaction rules, the latter gives an implicit representation of rules and allows a suitable description of combinatorial systems, e.g. protein complex formation. For both types of representations, there are tools for the incorporation and simulation of the dynamic behavior over a time-course are available (Fig. 8). The code-based analysis of systems represents a new class of analysis techniques, aiming to determine the possibility of a system to realize contingent molecular mappings, and to characterize them. It is a general-purpose method, but at the current state of research, restricted to explicit system models. It also does not consider dynamic properties of the system yet, as discussed above (Fig. 8, dashed lines). Nevertheless, it is useful in terms of giving certain predictions about the biological system, that afterwards are available for verification in wet-lab experiments.The algebraic approach (Fig. 8, red lines) could be based on abstract interpretation (Feret et al 2009), coarse-graining (Conzelmann et al 2006; Lenser et al 2007; Gruenert et al 2012), chemical organization theory (Kreyssig and Dittrich 2011; Kreyssig et al 2012; Speroni di Fenizio et al 2001), model checking (Chen et al 2013; Forejt et al 2012), or meaning space theory (De Boer and Verhoef 2012).

Outlook

In this paper we analyzed the effect of single CENPs on the probability for a molecular species to act in a certain code role. In general this approach can be extended by modeling combinations of CENPs, statistically spoken, by including the interaction terms in the GLMM. This can lead to new insights into the relation of CENPs to code roles, especially when an association is mediated mainly via a combination of CENPs (e.g. the STWX heterotetramer complex (Nishino et al 2012)). In such a case, the effect and significance of the main effects (i.e. the single Cenps) will decrease, while the interaction term gets significant. Technically the incorporation of all possible interaction is not feasible because of the large number of combinations of the 17 CENPs.

There are three points that need to be discussed regarding the type of data necessary for a code analysis. As discussed in (Görlich and Dittrich 2013) the code analysis is based on the assumption that all possible reactions are part of the reaction network model. Here, we tried to approximate the potential network by repeated simulation of the rule-based model. In general, it is also possible to approximate the potential network by merging different realizations, say in different species, or by knowledge-based approaches. Ideally, the used system model is complete, in a sense that all relevant molecular species and the reactions between these species are known. Although, modern biological and biochemical research help very much to increase our knowledge about the reaction networks of many biological subsystems, the current knowledge stays far from being complete, nor correct in some cases. The analysis of molecular codes depends on complete network information and thus results acquired are only valid in the context of the current model.

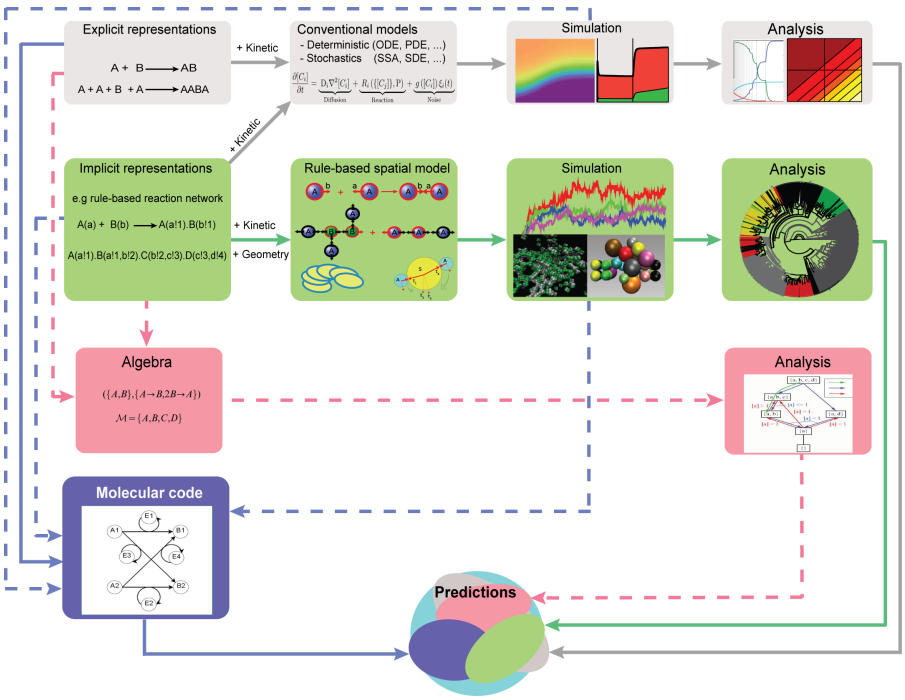

Fig. 8 Molecular code and computational modeling approaches
Illustration summarizes classical approaches (based on differential equations, in green), unconventional approaches (based on implicit rules, in gray), algebraic future approach (in red, dashed). Molecular code approach (this study, solid lilac line), and a future outlook of possible adaptations to the molecular code (in lilac, dashed line) ODE and PDE refers to the ordinary and partial differential equation. SSA refers to the stochastic simulation algorithm and SDE refers to the general stochastic differential equation which can be of any type (ordinary or partial).

The current algorithms for identifying molecular codes are based on the static network model of the respective system. Codes identified in such a network are not guaranteed to be "executable" in a dynamic setting (discussed above). Thus, in the future, dynamic information should be included into the code analysis and also energetic constraints can be considered (Savir and Tlusty 2013) in order to distinguish codes. Dynamic information can be obtained from either dynamic simulations of the system, e.g. via ordinary, or stochastic diffherential equations (cp. Fig. 8), or by direct measurement of time-course data of the respective system in te lab. Thus the incorporation of the time-axis does not only allow to validate identified codes, but also allows to find codes that can only be identified by incorporating the temporal scale.

The kinetochore is composed of several functional modules. The inner kinetochore, the outer kinetochore, which establishes the binding of the mitotic checkpoint (the Spindle Assembly Checkpoint, SAC) proteins. Thus, it would

be interesting to consider a full kinetochore and its participation in signaling and decision making.

References

Amaro AC, Samora CP, Holtackers R, Wang E, Kingston IJ, Alonso M, Lampson M, McAinsh AD, Meraldi P (2010) Molecular control of kinetochore-microtubule dynamics and chromosome oscillations. Nat Cell Biol 12:319–329

Barbieri M (2008a) Biosemiotics: a new understanding of life. Naturwissenschaften 95(7):577–599

Barbieri M (ed) (2008b) Introduction to Biosemiotics: The new Biological Synthesis. Springer, Dordrecht

Bergmann JH, Rodriguez MG, Martins NM, Kimura H, Kelly DA, Masumoto H, Larionov V, Jansen LE, Earnshaw WC (2011) Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore. EMBO J 30(2):328–340

Black BE, Cleveland DW (2011) Epigenetic centromere propagation and the nature of CENP-A nucleosomes. Cell 144:471–479

Caydasi AK, Lohel M, Gruenert G, Dittrich P, Pereira G, Ibrahim B (2012) Dynamical model of the Spindle Position Checkpoint. Molecular Systems Biology 8:582

Cheeseman IM, Desai A (2008) Molecular architecture of the kinetochore-microtubule interface. Nature Reviews Molecular Cell Biology 9(1):33–46

Chen T, Han T, Kwiatkowska M (2013) On the complexity of model checking interval-valued discrete time markov chains. Inform Process Lett

Cimini D, Degrassi F (2005) Aneuploidy: a matter of bad connections. Trends in Cell Biology 15(8):442 – 451

Conzelmann H, Saez-Rodriguez J, Sauter T, Kholodenko BN, Gilles ED (2006) A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks. BMC Bioinformatics 7:34

Dalal Y, Bui M (2010) Down the rabbit hole of centromere assembly and dynamics. Curr Opin Cell Biol 22:392–402

De Boer B, Verhoef T (2012) Language dynamics in structured form and meaning spaces. Advances in Complex Systems 15(3–4)

Doncic A, Ben-Jacob E, Barkai N (2005) Evaluating putative mechanisms of the mitotic spindle checkpoint. Proc Natl Acad Sci USA 102(18):6332–7

Faeder JR, Blinov ML, Hlavacek WS (2009) Rule-based modeling of biochemical systems with bionetgen. Methods in Molecular Biology 500:113–67

Farkas I, Derenyi I, Barabasi A, Vicsek T (2001) Spectra of "real-world" graphs: Beyond the semicircle law. Physical Review E 64(2):026,704

Favareau D (2010) Essential Readings in Biosemiotics. Springer, Dordrecht

Speroni di Fenizio P, Dittrich P, Banzhaf W (2001) Spontaneous formation of proto-cells in an universal artificial chemistry on a planar graph. In: Proceedings of the 6th European Conference on Advances in Artificial Life, Springer-Verlag, London, UK, UK, ECAL '01, pp 206–215

Feret J, Danos V, Krivine J, Harmer R, Fontana W (2009) Internal coarse-graining of molecular systems. Proc Natl Acad Sci USA 106(16):6453–6458

Foltz DR, Jansen LE, Black BE, Bailey AO, Yates JR, Cleveland DW (2006) The human CENP-A centromeric nucleosome-associated complex. Nat Cell Biol 8(5):458–469

Forejt V, Kwiatkowska M, Parker D (2012) Pareto curves for probabilistic model checking. In: Chakraborty S, Mukund M (eds) Proceeding of the 10th International Symposium on Automated Technology for Verification and Analysis, Springer, Lect. Notes Comp. Sci., vol 7561, pp 317–332

Gascoigne KE, Cheeseman IM (2010) Kinetochore assembly: if you build it, they will come. Current Opinion in Cell Biology In Press, Corrected Proof

Görlich D (2013) A formal model of molecular codes with respect to chemical reaction networks. PhD thesis, Friedrich-Schiller-Universität Jena

Görlich D, Dittrich P (2013) Molecular codes in biological and chemical reaction networks. PLoS ONE 8(1):e54,694, DOI 10.1371/journal.pone.0054694

Gruenert G, Ibrahim B, Lenser T, Lohel M, Hinze T, Dittrich P (2010) Rule-based spatial modeling with diffusing, geometrically constrained molecules. BMC Bioinformatics 11(1):307

Gruenert G, Escuela G, Dittrich P (2012) Symbol representations in evolving droplet computers. In: Proceedings of the 11th international conference on Unconventional Computation and Natural Computation, Springer-Verlag, Berlin, Heidelberg, UCNC'12, pp 130–140

Hanissian SH, Akbar U, Teng B, Janjetovic Z, Hoffmann A, Hitzler JK, Iscove N, Hamre K, Du X, Tong Y, Mukatira S, Robertson JH, Morris SW (2004) cDNA cloning and characterization of a novel gene encoding the MLF1-interacting protein MLF1IP. Oncogene 23(20):3700–3707

Harmer R (2010) Intrinsic information carriers in combinatorial dynamical systems. Chaos 20(3):037,108–

Holcombe M, Patton R (eds) (1998) Information Processing in Cells and Tissues. Plenum Press, New York

Holland AJ, Cleveland DW (2009) Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat Rev Mol Cell Biol 10(7):478–87

Howman EV, Fowler KJ, Newson AJ, Redward S, MacDonald AC, Kalitsis P, Choo KHA (2000) Early disruption of centromeric chromatin organization in centromere protein A (cenpa) null mice. Proc Natl Acad Sci USA 97(3):1148–1153

Hua S, Wang Z, Jiang K, Huang Y, Ward T, Zhao L, Dou Z, Yao X (2011) CENP-U cooperates with Hec1 to orchestrate kinetochore-microtubule attachment. J Biol Chem 286(2):1627–1638

Ibrahim B (2008) Systems Biology of Mitosis. Phd thesis, Friedrich-Schiller-University Jena

Ibrahim B, Dittrich P, Diekmann S, Schmitt E (2007) Stochastic effects in a compartmental model for mitotic checkpoint regulation. Journal of Integrative Bioinformatics 4(3):66

Ibrahim B, Diekmann S, Schmitt E, Dittrich P (2008a) In-silico modeling of the mitotic spindle assembly checkpoint. PLoS ONE 3(2):e1555

Ibrahim B, Dittrich P, Diekmann S, Schmitt E (2008b) Mad2 binding is not sufficient for complete cdc20 sequestering in mitotic transition control (an in silico study). Biophys Chem 134(1-2):93–100

Ibrahim B, Schmitt E, Dittrich P, Diekmann S (2009) In silico study of kinetochore control, amplification, and inhibition effects in MCC assembly. BioSystems 95:35–50

Klipp E, Wade RC, Kummer U (2010) Biochemical network-based drug-target prediction. Current Opinion in Biotechnology 21(4):511 – 516

Kreyssig P, Dittrich P (2011) Fragments and chemical organisations. Electr Notes Theor Comput Sci 272:19–41

Kreyssig P, Escuela G, Reynaert B, Veloz T, Ibrahim B, Dittrich P (2012) Cycles and the qualitative evolution of chemical systems. PLoS ONE 7(10):e45,772

Lenser T, Hinze T, Ibrahim B, Dittrich P (2007) Towards evolutionary network reconstruction tools for systems biology. In: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, 5th European Conference, pp 1500–1511

Li M, Fang X, Wei Z, York JP, Zhang P (2009) Loss of spindle assembly checkpoint-mediated inhibition of cdc20 promotes tumorigenesis in mice. J Cell Biol 185(6):983–94

Liu ST, Hittle JC, Jablonski SA, Campbell MS, Yoda K, Yen TJ (2003) Human CENP-I specifies localization of CENP-F, MAD1 and MAD2 to kinetochores and is essential for mitosis. Nat Cell Biol 5(4):341–345

Lohel M, Ibrahim B, Diekmann S, Dittrich P (2009) The role of localization in the operation of the mitotic spindle assembly checkpoint. Cell Cycle 8:2650–2660

Maiato H, DeLuca J, Salmon ED, Earnshaw WC (2004) The dynamic kinetochore-microtubule interface. J Cell Sci 117:5461–5477

Nishino T, Takeuchi K, Gascoigne KE, Suzuki A, Hori T, Oyama T, Morikawa K, Cheeseman IM, Fukagawa T (2012) CENP-T-W-S-X forms a unique centromeric chromatin structure with a histone-like fold. Cell 148:487–501

Obuse C, Yang H, Nozaki N, Goto S, Okazaki T, Yoda K (2004) Proteomics analysis of the centromere complex from HeLa interphase cells: UV-damaged DNA binding protein 1

(DDB-1) is a component of the CEN-complex, while BMI-1 is transiently co-localized with the centromeric region in interphase. Genes Cells 9(2):105–120

Okada M, Cheeseman IM, Hori T, Okawa K, McLeod IX, Yates JR, Desai A, Fukagawa T (2006) The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. Nat Cell Biol 8:446–457

Perpelescu M, Fukagawa T (2011) The abcs of cenps. Chromosoma 120:425–446

Quenet D, Dalal Y (2012) The CENP-A nucleosome: a dynamic structure and role at the centromere. Chromosome Res 20(5):465–479

Rohn H, Ibrahim B, Lenser T, Hinze T, Dittrich P (2008) Enhancing parameter estimation of biochemical networks by exponentially scaled search steps. In: Proceedings of the 6th European conference on Evolutionary computation, machine learning and data mining in bioinformatics, Springer-Verlag, Berlin, Heidelberg, EvoBIO'08, pp 177–187

Savir Y, Tlusty T (2013) The ribosome as an optimal decoder: A lesson in molecular recognition. Cell 153(2):471–479

Sebeok TA (2001) Biosemiotics: Its roots, proliferation, and prospects. Semiotica 134(1/4):61–78

Shannon CE (1948) A mathematical theory of communication. Bell system technical journal 27

Stoler S, Keith KC, Curnick KE, Fitzgerald-Hayes M (1995) A mutation in CSE4, an essential gene encoding a novel chromatin-associated protein in yeast, causes chromosome nondisjunction and cell cycle arrest at mitosis. Genes Dev 9(5):573–586

Suijkerbuijk SJ, Kops GJ (2008) Preventing aneuploidy: the contribution of mitotic checkpoint proteins. Biochim Biophys Acta 1786(1):24–31

Sullivan KF, Hechenberger M, Masri K (1994) Human cenp-A contains a histone h3 related histone fold domain that is required for targeting to the centromere. The Journal of Cell Biology 127(3):581–592

Topp CN, Zhong CX, Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci USA 101(45):15,986–15,991

Tschernyschkow S, Herda S, Gruenert G, Doring V, Gorlich D, Hofmeister A, Hoischen C, Dittrich P, Diekmann S, Ibrahim B (2013) Rule-based Modeling and Simulations of the Inner Kinetochore Structure. Prog Biophys Mol Biol In Press

Waltermann C, Klipp E (2011) Information theory based approaches to cellular signaling. Biochim Biophys Acta 1810:924–932

Appendix

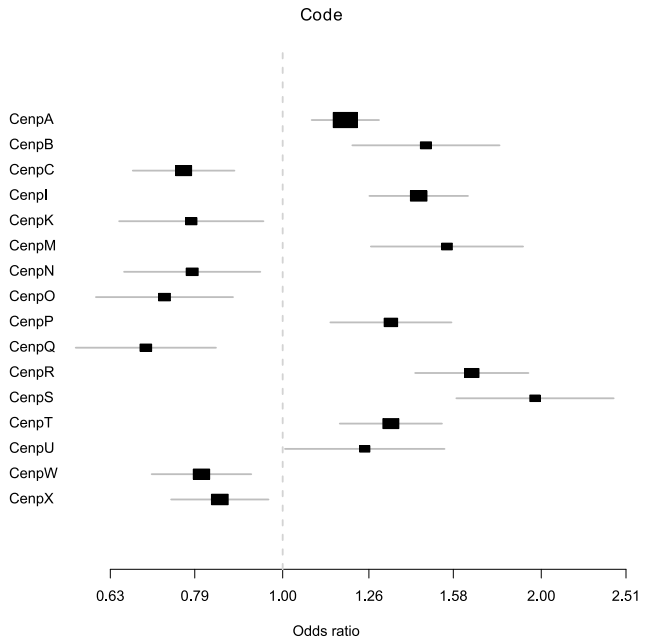

Fig. 9 Forest plot of the odds ratios and 95%-confidence intervals of the GLMM for response CODE

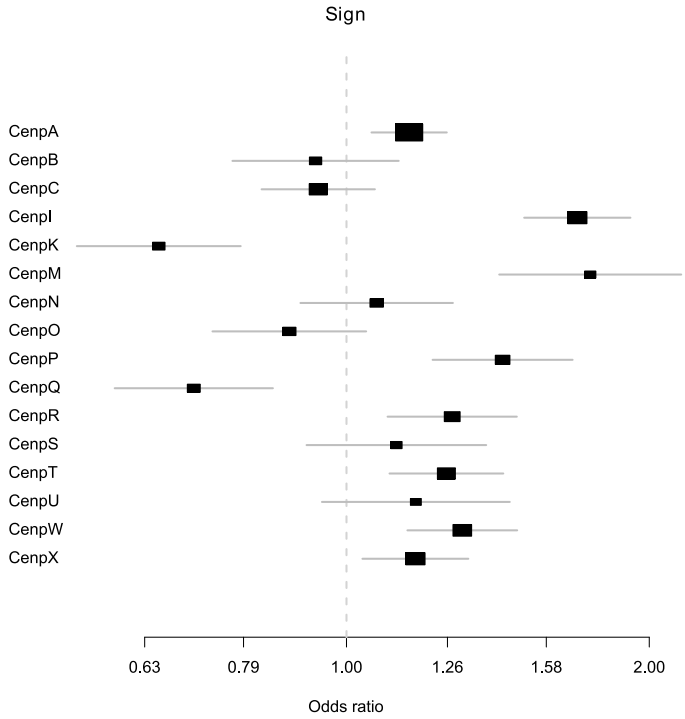

Fig. 10 Forest plot of the odds ratios and 95%-confidence intervals of the GLMM for response SIGN

Table 4 Odd ratio estimates of the generalized linear mixed model for the response SIGN.

						95%-CI	
	Protein	Estimate	StdErr	Significance	OR	Lower	Upper
1	**CenpA**	0.14	0.04	**0.0011**	**1.15**	**1.06**	**1.26**
2	CenpB	-0.07	0.10	0.4660	0.93	0.77	1.13
3	CenpC	-0.06	0.07	0.3285	0.94	0.82	1.07
4	**CenpI**	0.53	0.06	**<.0001**	**1.70**	**1.50**	**1.92**
5	**CenpK**	-0.43	0.10	**<.0001**	**0.65**	**0.54**	**0.78**
6	**CenpM**	0.56	0.11	**<.0001**	**1.75**	**1.42**	**2.15**
7	CenpN	0.07	0.09	0.4381	1.07	0.90	1.28
8	CenpO	-0.13	0.09	0.1447	0.88	0.74	1.05
9	**CenpP**	0.36	0.08	**<.0001**	**1.43**	**1.22**	**1.68**
10	**CenpQ**	-0.35	0.09	**0.0002**	**0.70**	**0.59**	**0.85**
11	**CenpR**	0.24	0.08	**0.0014**	**1.27**	**1.10**	**1.48**
12	CenpS	0.11	0.10	0.2766	1.12	0.91	1.38
13	**CenpT**	0.23	0.07	**0.0006**	**1.26**	**1.10**	**1.43**
14	CenpU	0.16	0.11	0.1476	1.17	0.95	1.45
15	**CenpW**	0.27	0.06	**<.0001**	**1.30**	**1.15**	**1.48**
16	**CenpX**	0.16	0.06	**0.0108**	**1.17**	**1.04**	**1.32**

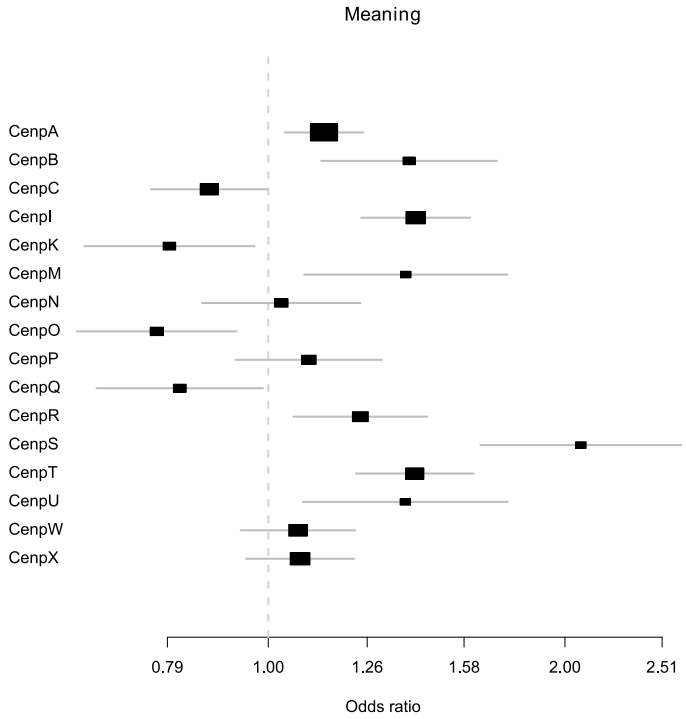

Fig. 11 Forest plot of the odds ratios and 95%-confidence intervals of the GLMM for response MEANING

Table 5 Odd ratio estimate of the generalized linear mixed model for the response MEANING.

						95%-CI	
	Protein	Estimate	StdErr	Significance	OR	Lower	Upper
1	**CenpA**	0.13	0.05	**0.0054**	**1.14**	**1.04**	**1.25**
2	**CenpB**	0.33	0.10	**0.0017**	**1.39**	**1.13**	1.25
3	**CenpC**	-0.14	0.07	**0.0485**	**0.87**	**0.76**	**1.00**
4	**CenpI**	0.34	0.06	**<.0001**	**1.41**	**1.24**	**1.60**
5	CenpK	-0.23	0.10	**0.0230**	**0.79**	**0.65**	**0.97**
6	**CenpM**	0.32	0.12	**0.0081**	**1.38**	**1.09**	**1.75**
7	CenpN	0.03	0.09	0.7465	1.03	0.86	1.24
8	**CenpO**	-0.26	0.10	**0.0064**	**0.77**	**0.64**	**0.93**
9	CenpP	0.09	0.09	0.2763	1.10	0.93	1.30
10	CenpQ	-0.21	0.10	**0.0373**	**0.81**	**0.67**	**0.99**
11	**CenpR**	0.22	0.08	**0.0071**	**1.24**	**1.06**	**1.45**
12	**CenpS**	0.73	0.12	**<.0001**	**2.08**	**1.64**	**2.62**
13	**CenpT**	0.34	0.07	**<.0001**	**1.41**	**1.23**	**1.62**
14	**CenpU**	0.32	0.12	**0.0088**	**1.38**	**1.08**	**1.75**
15	CenpW	0.07	0.07	0.3094	1.07	0.94	1.23
16	CenpX	0.07	0.06	0.2480	1.08	0.95	1.22

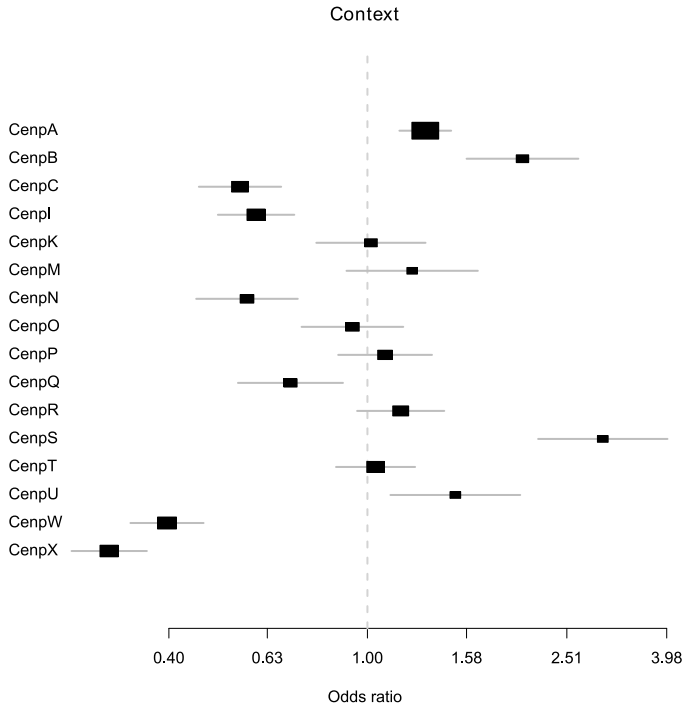

Fig. 12 Forest plot of the odds ratios and 95%-confidence intervals of the GLMM for response CONTEXT.

Table 6 Odd ratio estimates of the generalized linear mixed model for the response CONTEXT.

						95%-CI	
	Protein	Estimate	StdErr	Significance	OR	Lower	Upper
1	**CenpA**	0.27	0.06	**<.0001**	**1.31**	**1.16**	**1.47**
2	**CenpB**	0.72	0.13	**<.0001**	**2.05**	**1.58**	**2.65**
3	**CenpC**	-0.59	0.10	**<.0001**	**0.56**	**0.46**	**0.67**
4	**CenpI**	-0.51	0.09	**<.0001**	**0.60**	**0.50**	**0.71**
5	CenpK	0.02	0.13	0.8998	1.02	0.79	1.31
6	CenpM	0.21	0.15	0.1818	1.23	0.91	1.66
7	**CenpN**	-0.56	0.12	**<.0001**	**0.57**	**0.45**	**0.73**
8	CenpO	-0.07	0.12	0.5628	0.93	0.74	1.18
9	CenpP	0.08	0.11	0.4603	1.08	0.87	1.35
10	**CenpQ**	-0.36	0.12	**0.0041**	**0.70**	**0.55**	**0.89**
11	CenpR	0.15	0.10	0.1346	1.17	0.95	1.43
12	**CenpS**	1.09	0.15	**<.0001**	**2.96**	**2.20**	**3.99**
13	CenpT	0.04	0.09	0.6868	1.04	0.86	1.25
14	**CenpU**	0.41	0.15	**0.0079**	**1.50**	**1.11**	**2.02**
15	**CenpW**	-0.92	0.09	**<.0001**	**0.40**	**0.34**	**0.47**
16	**CenpX**	-1.19	0.09	**<.0001**	**0.30**	**0.26**	**0.36**