

6 Sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The NtrY/HIG manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript „*Sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins*“ which is submitted for publication in *Cellular Signalling*, declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub

- conducted the protein sequence analysis,
- discovered the similarity between NtrY and HIG protein families,
- wrote the text and prepared the figures for the manuscript,
- serves as the corresponding author during the review process.

2) Thomas Braun

- raised the interest in the search for vertebrate homologs of bacterial proteins that function in histidine phosphorylation-dependent signalling,
- contributed to the manuscript preparation by hints to relevant literature and by comments on the style of the manuscript.

Sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins

Eike Staub^{§*} and Thomas Braun[#]

[§] metaGen Pharmaceuticals GmbH, Oudenarder Str. 16, D-13347 Berlin, Germany

current address: Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology,
Ihnestr. 73, D-14195 Berlin, Germany

[#] University of Halle-Wittenberg, Institute of Physiological Chemistry, Hollystr. 1, D-06097 Halle, Germany

* author to whom correspondence should be addressed

email: staub@molgen.mpg.de

tel ++49-(0)30-8413-1157

fax ++49-(0)30-8413-1152

Abstract

Whereas in eukaryotic signalling pathways protein kinases mainly act on serine, threonine or tyrosine residues, in bacterial signal transduction predominantly histidine residues become phosphorylated. Although it is known that histidine phosphorylation also occurs in eukaryotes, the players and mechanisms of metazoan histidine phosphorylation remain elusive. Here we demonstrate that proteins encoded by the hypoxia-inducible gene (HIG) family, a group of mammalian proteins of unknown biochemical function, have significant sequence similarity to a subfamily of prokaryotic histidine kinases, which function in the response to nitrogen. The similarity region comprises the typical membrane-proximal sensory domains of eubacterial NtrY-like histidine kinases and the membrane-proximal regions of plant, fungal and metazoan HIG proteins. To our knowledge, this is the first report of significant sequence similarity which links a sensory domain of a prokaryotic two-component signal transduction pathway to a family of vertebrate proteins. Based on this sequence similarity and on a potential functional link of both families, the involvement in cellular processes dependent on the levels of gaseous compounds, we hypothesise that HIG and NtrY protein families are homologous. We suggest that HIG proteins are top candidates for future experimental studies that try to link bacterial phosphohistidine-dependent signal transduction to metazoan cellular signalling.

Introduction

Protein phosphorylation is a common signal transduction mechanism in all organisms. In eukaryotes the most widely known protein kinases catalyse the phosphorylation of hydroxyamino acids, i.e. serine/threonine protein kinases and tyrosine kinases. In contrast, the phosphorylation of histidine residues, which is the predominant type of protein phosphorylation in bacterial signal transduction, is only poorly characterised in mammals. Protein kinases specific for histidines have not yet been described in vertebrates, only in plants, fungi and prokaryotes ^(1,2). Only rough estimates about the degree of histidine phosphorylation in mammals exists ⁽³⁻⁵⁾. In lower eukaryotes, however, 6% of the phosphoamino acids of basic nuclear proteins are phosphohistidines, which is about two orders of magnitudes greater than the abundance of phosphotyrosine in this subset of proteins ^(6,7). This suggests that the importance of histidine phosphorylation in eukaryotic signal transduction has not been fully recognised yet.

In prokaryotes, histidine phosphorylation is the key mechanism of so called 'two-component' signalling pathways, which link extracellular stimuli such as changing osmolarity, oxygen, or nitrogen levels to gene regulation, but also affect other functions. In two-component pathways, signals are sensed by cell surface-located receptors, the sensors. This results in their dimerisation and in auto-phosphorylation of cytoplasmic histidines. The high energy histidine-bound phosphates are transferred to aspartates in the receiver domains of response regulators. These proteins are often transcriptional regulators, which are activated upon phosphorylation-mediated conformational changes ⁽⁸⁾. Typically, receptor histidine kinases have an extracellular sensory loop flanked by transmembrane regions, a dimerisation domain and a kinase domain ⁽⁸⁾.

One bacterial subfamily of histidine kinases is composed of the NtrY receptors, which are involved in the control of nitrogen-related environmental stimuli ⁽⁹⁾. Recently, the hypoxia-inducible gene (HIG) family has been identified in vertebrates. Protein products of this family have not yet been characterised biochemically. The founding gene HIG is upregulated in response to hypoxia in the hypoxia-tolerant fish *Gillichthys mirabilis* ⁽¹⁰⁾. Both families are predicted to comprise α -helical transmembrane proteins of largely uncharacterised biochemical function. During this study the sequence similarity between these two families is investigated. The results are discussed in the light of their potential meaning for cellular signalling in metazoa.

Materials and Methods

For all types of sequence similarity searches in databases we used the non-redundant protein database (nr) at the NCBI (<http://www.ncbi.nlm.nih.gov/Database/>) and the EBI set of bacterial protein sequence databases derived from completely sequenced bacterial genomes (<http://www.ebi.ac.uk/protomes/>). An initial HMM of the HIG protein family (HIG_1_N) was obtained from the Pfam database (version 7.0) of protein families ⁽¹¹⁾. Searches for local sequence similarity between query protein sequences and database proteins were carried out using BLASTP and PSIBLAST using three substitution matrices (PAM250, BLOSUM62, BLOSUM45) in combination with various E value cut-offs and profile inclusion thresholds ⁽¹²⁾. Protein alignments were constructed using CLUSTALX ⁽¹³⁾. HMM models of protein alignments were built and calibrated using the HMMER package ⁽¹⁴⁾. The E value statistic of each HMM was calibrated using the default options of hmmscalibrate; the scores of 5000 random sequences, each of 325 residues length, were fitted to an extreme value distribution that was subsequently used for the calculation of E values for query scores. The PRSS program of the FASTA package was used to align two pairs of sequences and to assign a P value as an estimate for the significance of the alignments. The P value is estimated on the distribution of scores from alignments of one sequence with a randomly shuffled version of the other sequence ⁽¹⁵⁾. To obtain an estimate for the similarity between two alignments, we used the LAMA web server with default options (minimal alignment length of 4 residues, minimum reported Z score of 5.6, calculation of the expectation (E) value on the basis of 5000 blocks (1700 more than in version 9.1 of the BLOCKS database) ⁽¹⁶⁾. Using the COMPASS program we identified local similarities between alignment profiles using a Smith-Waterman-like algorithm allowing for the insertion of gapped columns during profile-to-profile alignment ⁽¹⁷⁾. The calculation of E values for the resulting profile-to-profile alignments is based on the number of aligned columns in a profile database that can be specified explicitly.

Results and Discussion

The experimental evidence for a function of HIG proteins in hypoxia and the presence of two transmembrane domains in the N-terminal region of HIG proteins, like in histidine kinases, encouraged us to investigate the role of these proteins by sequence analysis. Using a Hidden Markov Model (HMM) of the HIG family from the Pfam database we scanned the protein databases of several eubacteria for HIG homologues. Indeed, we found a marginal similarity of HIG proteins to a number of bacterial proteins, among these the NtrY protein of *Caulobacter crescentus*. Using

this sequence fragment as a query, we performed PSIBLAST searches in the non-redundant (nr) protein database of the NCBI using various cut-offs. Applying an E value inclusion threshold of 0.01 we identified a family of 19 NtrY-like proteins. However, we were not able to detect significant similarity to HIG proteins by reciprocal PSIBLAST searches.

To gain sensitivity in database searches, we built an alignment of the putative sensory transmembrane region of NtrY proteins and derived a profile HMM by the use of the hmmbuild program of the HMMER package. The application of such an HMM on the HIG proteins resulted in alignments with the NtrY model consensus that hardly showed gapped regions. Two hits had E values of 0.033 and 0.05. For reciprocal HMM analysis, we built a HMM from a subsection of the Pfam HIG alignment (name HIG_1_N) which comprised the putative NtrY homology region. The application of the HIG HMM to NtrY-like proteins resulted in four hits with E values less than 0.06. However, because in searches of the large nr database using these HMMs the E values were no longer significant, these findings cannot be regarded as a proof of significant similarity.

To confirm the marginal similarity found in reciprocal HMM searches by an independent method, we applied a complementary approach based on the extensive pairwise cross-comparison of single sequences from one subfamily with single sequences from the other subfamily using the PRSS program. For the resulting 154 Smith-Waterman alignments, the median P value of all comparisons was 0.25, the 25th percentile was 0.09, and the minimum P value was 0.0022. By this analysis we showed that not only the composition, but also the order of amino acid residues in sequences of both families contributes to the similarity. Because multiple sequence pairs of the two families can be aligned with significant P values below 0.05 we argue that this is further evidence for the significance of inter-family similarity.

To further investigate the hypothesis of an ancestral relation between the two families we applied two profile-to-profile alignment comparison methods. The LAMA method identifies ungapped homologous BLOCKS between two protein sequence alignments and estimates E values for the findings. The alignments of the HIG and NtrY proteins both passed the check for biased composition of BLOCKS. LAMA found a common BLOCK of 50 residues length with a score of 24. Assuming a database size of 5000 Blocks, this results in a significant Z score of 7.1 and an E value of 1.5e-2. This is further evidence that the sequence similarity between the two protein families is significant.

Because LAMA does not allow for gaps in the identified common BLOCK of two alignments, it might lose sensitivity. Therefore, we applied a second profile-to-profile comparison method allowing for gaps. The COMPASS method identifies local similarities between alignment profiles using a Smith-Waterman-like algorithm. The

calculation of E values for the resulting profile-to-profile alignments is based on the number of aligned columns in a profile database. When we compared the two automatic CLUSTALX alignments of the complete sequences of the NtrY and HIG protein sets, COMPASS identified a common region that is nearly identical to the aligned region shown in figure 1 with a score of 108 and an E value of $1.25e-10$ when the database size is set to the length of the larger alignment. Explicitly specifying a database size of 1.000.000, which is greater than the number of aligned columns with gap fraction < 0.5 in the Pfam database, resulted in an E value of $3.08e-6$. The COMPASS results clearly indicate significant sequence similarity in the common region of NtrY and HIG proteins. This is further evidence that both families are homologous. We argue that this is not due to a bias in amino acid composition that could arise from the incorporation of transmembrane helix residues, because (a) the PRSS analysis showed that the order of amino acids in HIG and NtrY sequences contributes significantly to the quality of pairwise sequence alignments and (b) the alignments of individual families passed the check for composition-biased alignments as implemented in the LAMA method. In the further course of the analysis we assume that NtrY and HIG proteins are homologous.

We constructed a combined alignment of the common region of the two subfamilies. Using the profile-to-profile alignment mode of CLUSTALX we obtained a combined NtrY/HIG superfamily alignment. We iteratively constructed a HMM, scanned the nr database for further homologues, and built a new HMM. The iterative searches converged in the 3rd round, resulting in the identification of 62 sequences in the nr database. After removal of redundancy at the 95% identity level, 48 sequences remained (see alignment figure 1). Among those were sequences from yeast, a filamentous fungus, plants, fly, mosquito and worm. This means that members of the postulated NtrY/HIG superfamily of proteins are present in all phyla, with the notable exception of the archaeobacterial lineage. We hypothesise that the eukaryotic NtrY/HIG-like proteins are a eubacterial invention.

Furthermore, we argue that members of the HIG protein family, of which one member is upregulated in response to hypoxia at the transcript level, are probably involved in the sensing of small molecules. This might include sensing of local concentrations of oxygen, nitrogen, or NO near the surface of an animal cell. It is not clear from our work which specific member of the superfamily might detect what small molecule. However, it is tempting to speculate that HIG proteins sense oxygen and NtrY proteins sense nitrogen or nitrogen-related gaseous compounds taking into consideration the context in which these proteins were discovered. The alignment predicts a special role for basic residues in the central region of the domain (see residues 30 and 31 in figure 1). Their negative charges could be important for the binding of anions or gaseous compounds with partially negative

charge. The hydrophobicity in the transmembrane region is highly conserved, especially the aliphatic residues 13, 39 and 52 and a single tiny hydrophobic amino acid in residue 54, suggesting that these residues are important for the structure of the receptors.

A throughout evaluation of the domain architecture of NtrY and HIG proteins revealed that the vast majority of NtrY-like proteins have typical domains of histidine kinase receptors like HAMP domains, PAS domains, phosphoacceptor domains and the kinase domains themselves. Of the two NtrY-like proteins that lack these domains, one is described as a fragmentary sequence and the other is a hypothetical protein. The HIG-like proteins are devoid of such domains and have much shorter cytoplasmic tails. Two hypothetical HIG-like proteins with unusual domain composition stood out. The predicted rat protein XP_228571.1 most likely resulted from an erroneous gene prediction, leading to a fusion of a ribosomal L18ae-like protein with a HIG-like domain. In the hypothetical *Arabidopsis thaliana* T17F15.100 protein a RING finger domain was fused to the C-terminal part. RING fingers are known to be the catalytic domains of E3 ubiquitin ligases that confer the target specificity to ubiquitin-dependent protein degradation. Like the sequence itself, this functional link to proteasomal protein degradation is hypothetical. It remains an open question whether metazoan proteins have lost their additional histidine kinase domains or the bacterial sequences gained them. Since almost all bacterial NtrY-like proteins share the typical histidine kinase domain, we reason that the postulated common ancestor of HIG/NtrY proteins was a histidine kinase and that early domain loss in the metazoan lineage shaped the contemporary HIG proteins.

In conclusion, we have identified a eukaryotic protein domain in a broad range of species including vertebrates that has significant similarity to bacterial sensory domains of histidine kinase receptors. There is only limited knowledge about the biochemical functions of both families. Both seem to be involved in processes in which the concentration of gases, here nitrogen and oxygen, play a role. This weak functional link is in accordance with our assumption of homology. A further proof that the similarity between both families is not due to convergent evolution of unrelated sequences, but instead is due to evolution by descent from a common ancestral sequence, has to come from future experimental studies. Independent of the question whether the observed similarity is due to analogy or homology, we expect that the biochemical analysis of mammalian HIG proteins will lead to new insights into the mechanisms which allow animal cells to sense small compounds like gases in their local environment. Based on the assumption of homology we suggest that the HIG family of proteins is a good starting point for studies that try to discover the source of phosphohistidine-dependent signal transduction in

mammalian cells. However, even if HIG proteins will be confirmed as the first vertebrate homologs of histidine kinases, the source of histidine phosphorylation in eukaryotes would still remain unknown, because HIG proteins do not have a cytoplasmic histidine kinase domain. Nevertheless, HIG proteins as possible sensory receptors could then facilitate the search for downstream phosphohistidine signalling proteins by biochemical means. Therefore, we consider the HIG-like proteins to be important molecules for the elucidation of mechanisms of histidine phosphorylation in eukaryotic proteins.

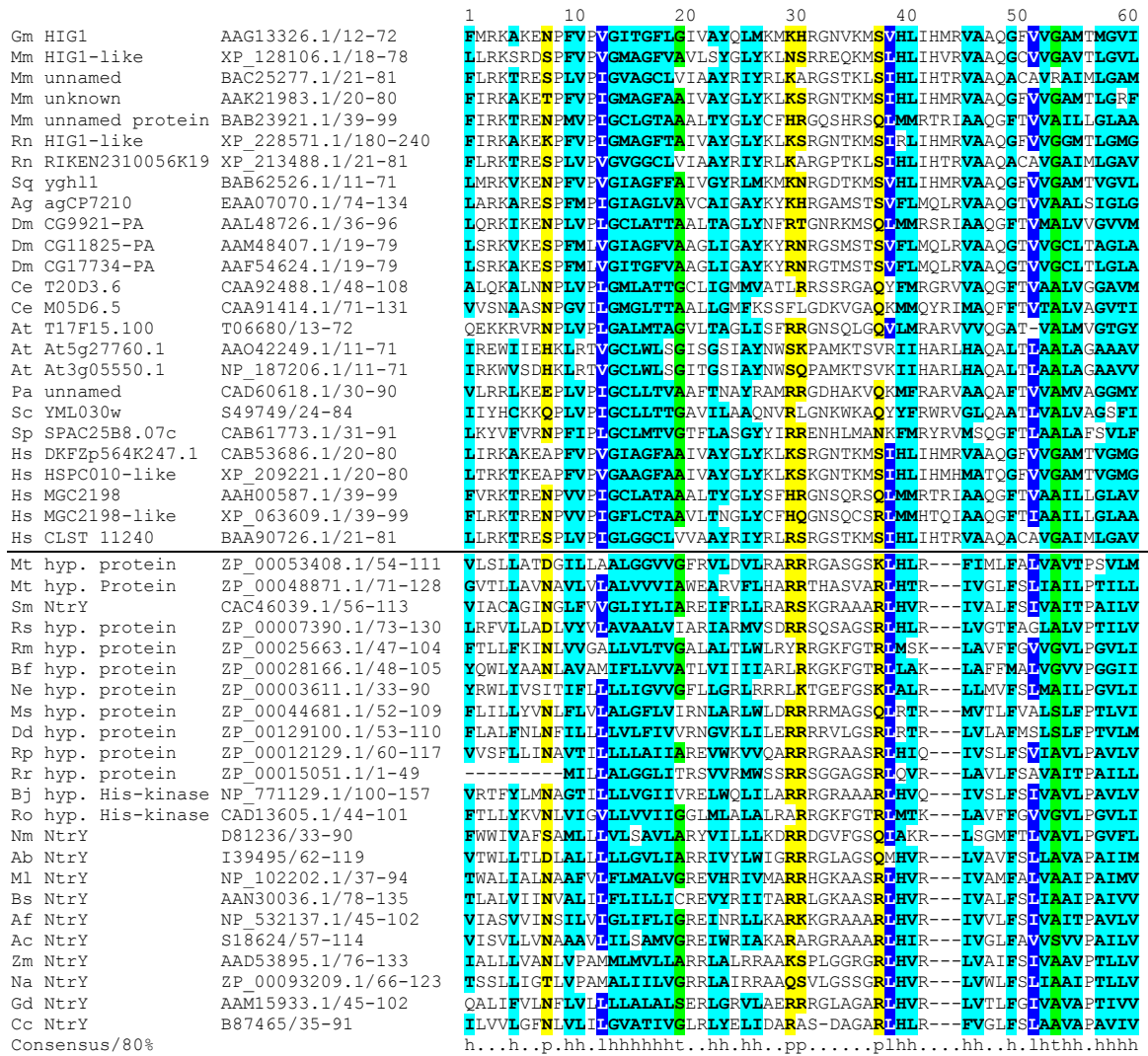


Figure 1

Sequence alignment of the common domain of NtrY and HIG proteins. Each line contains a two-letter organism code, the protein name, a sequence database identifier (NCBI non-redundant protein database) followed by the location of the displayed sequence fragment and the domain sequence itself. Organism code: Hs *Homo sapiens*, Mm *Mus musculus*, Rn *Rattus norvegicus*, Sq *Seriola quinqueradiata*, Ag *Anopheles gambiae*, Dm *Drosophila melanogaster*, Ce *Caenorhabditis elegans*, At *Arabidopsis thaliana*, Pa *Podospira anserine*, Sc *Saccharomyces cerevisiae*, Sp *Schizosaccharomyces pombe*, Rs *Rhodobacter sphaeroides*, Rm *Ralstonia metallidurans*, Ro *Ralstonia solanacearum*, Bf *Burkholderia fungorum*, Ne *Nitrosomonas europaea*, Nm *Neisseria meningitidis* MC58, Mt *Magnetospirillum magnetotacticum*, Ms *Magnetococcus* sp. MC-1, Dd *Desulfovibrio desulfuricans* G20, Ab *Azospirillum brasilense*, Ml *Mesorhizobium loti*, Bs *Brucella suis* 1330, Sm *Sinorhizobium meliloti*, Af *Agrobacterium tumefaciens*, Bj *Bradyrhizobium japonicum*, Rp *Rhodospseudomonas palustris*, Ac *Azorhizobium caulinodans*, Rr *Rhodospirillum rubrum*, Zm *Zymomonas mobilis*, Na *Novosphingobium aromaticivorans*, Gd *Gluconacetobacter diazotrophicus*, Cc *Caulobacter crescentus* CB15. The amino acids are coloured according an 80% consensus rule and the following classification: DE negative (-) yellow, ST hydroxy (*) brown, ILV aliphatic (l) dark blue, HKR positive (+) red, AGS tiny (t) green, FHWY aromatic (a) purple, DEHKR charged (c) dark green, CDEHKNQIRST polar (p) light orange, ACFGHILMTVWY hydrophobic (h) light blue. The horizontal bar separates the HIG subfamily and the NtrY subfamily.

References

1. Hwang I, Chen HC, Sheen J. Two-component signal transduction pathways in *Arabidopsis*. *Plant Physiol* 2002;129(2):500-515.
2. Santos JL, Shiozaki K. Fungal histidine kinases. *Sci STKE* 2001;2001(98):RE1.
3. Chen CC, Bruegger BB, Kern CW, Lin YC, Halpern RM, Smith RA. Phosphorylation of nuclear proteins in rat regenerating liver. *Biochemistry* 1977;16(22):4852-4855.
4. Chen CC, Smith DL, Bruegger BB, Halpern RM, Smith RA. Occurrence and distribution of acid-labile histone phosphates in regenerating rat liver. *Biochemistry* 1974;13(18):3785-3789.
5. Smith DL, Bruegger BB, Halpern RM, Smith RA. New histone kinases in nuclei of rat tissues. *Nature* 1973;246(5428):103-104.
6. Matthews HR, Huebner VD. Nuclear protein kinases. *Mol Cell Biochem* 1984;59(1-2):81-99.
7. Matthews HR. Protein kinases and phosphatases that act on histidine, lysine, or arginine residues in eukaryotic proteins: a possible regulator of the mitogen-activated protein kinase cascade. *Pharmacol Ther* 1995;67(3):323-350.
8. Wolanin PM, Thomason PA, Stock JB. Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome Biol* 2002;3(10):REVIEWS3013.
9. Pawlowski K, Klosse U, de Bruijn FJ. Characterization of a novel *Azorhizobium caulinodans* ORS571 two- component regulatory system, NtrY/NtrX, involved in nitrogen fixation and metabolism. *Mol Gen Genet* 1991;231(1):124-138.
10. Gracey AY, Troll JV, Somero GN. Hypoxia-induced gene expression profiling in the euryoxic fish *Gillichthys mirabilis*. *Proc Natl Acad Sci U S A* 2001;98(4):1993-1998.
11. Bateman A, Birney E, Cerruti L, Durbin R, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res* 2002;30(1):276-280.
12. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 2001;29(14):2994-3005.
13. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* 1998;23(10):403-405.
14. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;14(9):755-763.
15. Pearson WR. Effective protein sequence comparison. *Methods Enzymol* 1996;266:227-258.
16. Pietrokovski S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* 1996;24(19):3836-3845.
17. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* 2003;326(1):317-336.