2011

Use of comparative genomics approaches to characterize interspecies differences in response to environmental chemicals: Challenges, opportunities, and research needs

Sarah L. Burgess-Herbert
U.S. EPA, sarah.burgess@alum.mit.edu

Susan Y. Euling
U.S. EPA

Toxicology and Applied Pharmacology xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Toxicology and Applied Pharmacology

journal homepage: www.elsevier.com/locate/ytaap

Contemporary Issues in Toxicology

Use of comparative genomics approaches to characterize interspecies differences in response to environmental chemicals: Challenges, opportunities, and research needs ☆

Sarah L. Burgess-Herbert [a,*], Susan Y. Euling [b]

[a] American Association for the Advancement of Science (AAAS) Science and Technology Policy Fellow at the US Environmental Protection Agency (EPA), 2009–10, USA
[b] National Center for Environmental Assessment, Office of Research and Development, US Environmental Protection Agency, Washington, DC 20460, USA

ARTICLE INFO

Article history:
Received 8 June 2011
Revised 11 November 2011
Accepted 16 November 2011
Available online xxxx

Keywords:
Cross-species
Molecular network
Biological pathway
Selective constraints
-omics
Human health risk assessment
Systems biology

ABSTRACT

A critical challenge for environmental chemical risk assessment is the characterization and reduction of uncertainties introduced when extrapolating inferences from one species to another. The purpose of this article is to explore the challenges, opportunities, and research needs surrounding the issue of how genomics data and computational and systems level approaches can be applied to inform differences in response to environmental chemical exposure across species. We propose that the data, tools, and evolutionary framework of comparative genomics be adapted to inform interspecies differences in chemical mechanisms of action. We compare and contrast existing approaches, from disciplines as varied as evolutionary biology, systems biology, mathematics, and computer science, that can be used, modified, and combined in new ways to discover and characterize interspecies differences in chemical mechanism of action which, in turn, can be explored for application to risk assessment. We consider how genetic, protein, pathway, and network information can be interrogated from an evolutionary biology perspective to effectively characterize variations in biological processes of toxicological relevance among organisms. We conclude that comparative genomics approaches show promise for characterizing interspecies differences in mechanisms of action, and further, for improving our understanding of the uncertainties inherent in extrapolating inferences across species in both ecological and human health risk assessment. To achieve long-term relevance and consistent use in environmental chemical risk assessment, improved bioinformatics tools, computational methods robust to data gaps, and quantitative approaches for conducting extrapolations across species are critically needed. Specific areas ripe for research to address these needs are recommended.

Contents

Abbreviations: AhR, aryl hydrocarbon receptor; ADME, absorption, **d**istribution, **m**etabolism, and **e**xcretion; APAP, N-acetyl-*p*-aminophenol; EGBE, ethylene glycol monobutyl ether; MYC, myelocytomatosis oncogene; MOA, mode of action; DRE, dioxin response element; HMGCS2, 3-hydroxy-3-methylglutaryl-CoA synthase 2 (mitochondrial); IRIS, Integrated Risk Information System; JUN, jun proto-oncogene; PPARα, peroxisome proliferator activated receptor alpha; RfC, reference concentration; RfD, reference dose; NOAEL, no-observed-adverse-effect-level; LOAEL, lowest-observed-adverse-effect-level; DBP, dibutyl phthalate; TCDD, 2,3,7,8-tetrachlorodibenzo-*p*-dioxin; TD, toxicodynamic; TK, toxicokinetic; NRC, National Research Council; EPA, Environmental Protection Agency; DNA, deoxyribonucleic acid; RNA, ribonucleic acid; PID, percent identity; PXR, pregnane X receptor.

☆ Disclaimer: This manuscript has been reviewed by the U.S. Environmental Protection Agency and approved for publication. The views expressed in this manuscript are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

* Corresponding author at: 2151 Jamieson Ave #1504, Alexandria, VA, 22314, USA. Tel.: +1 703 624 1989.
E-mail address: sarah.burgess@alum.mit.edu (S.L. Burgess-Herbert).

0041-008X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.taap.2011.11.011

Introduction

In the National Research Council's (NRC) 2007 report, Toxicity Testing in the 21st Century: A Vision and a Strategy, they provided a general plan for a transition to a new "pathway-based" paradigm for toxicity testing and risk assessment (National Research Council, 2007). This strategy envisions the replacement of *in vivo* animal toxicology studies with *in vitro* molecular studies, and the use of these *in vitro* studies in combination with computational models as the basis for regulatory decision making. In response, the U.S. Environmental Protection Agency (EPA) has developed its own strategic plan for moving towards a toxicity pathway-based risk assessment paradigm (U.S. EPA, 2009; www.epa.gov/spc/toxicitytesting). Genomics data, resources, and tools are central to the execution of this future risk assessment paradigm. In addition, they are also useful within the current paradigm of risk assessment, especially in the context of formulating evidence-based extrapolations from one population to another and from one species to another. Indeed, one of the three key areas identified by a Health and Environmental Sciences Institute (HESI) survey in which genomics is perceived as having the greatest relevance and impact on toxicology is in the identification of species differences (Pettit et al., 2010). Identifying and characterizing these interspecies differences is requisite to better informing the extrapolation of chemical risk across species, including from test animals to humans, and to understanding the uncertainties involved in this process.

There can be important differences among species in their response to chemical exposure, making the extrapolation of toxicological inferences from animal models to humans challenging in the context of human health risk assessment, and inferences across species challenging in the context of ecological risk assessment. These differences can be in dosimetry, exposure, metabolic pathways, and in homology of genes, proteins, biochemistry, and physiology (National Research Council, 2006). Because of these challenges, there is a need for approaches to identify, understand, and, ideally, quantify interspecies differences in chemical mechanism of action. A chemical's mechanism of action is defined here as the complete molecular sequence of events between the interaction of the chemical with its target site(s) and observation of the outcome(s) (Fig. 1). The mechanism of action can include both toxicokinetic (e.g., chemical absorption, distribution, metabolism, or excretion) and toxicodynamic events (e.g., changes at the molecular, cellular, tissue, or organ level). Because a chemical's mechanism of action has rarely been fully understood due to data limitations, the "mode of action" concept was developed for and applied to human health risk assessment. A chemical's mode of action, or "MOA," has been functionally defined as a key event, or a sequence of key events, that a particular toxicity outcome is dependent upon (i.e., part of the causal pathway and not a coincident event) (U.S. EPA, 2006). However, increasing information on mechanism of action provided by genomic and toxicogenomic studies (for example, see the Comparative Toxicogenomics Database; ctd.mdibl.org) has revealed that numerous environmental chemicals have multiple MOAs, in some cases affecting multiple pathways. Therefore, we focus here on the types of genomics data, resources, and associated computational tools that show promise for characterizing interspecies differences in mechanisms of action, and that will thereby improve our understanding of the uncertainties

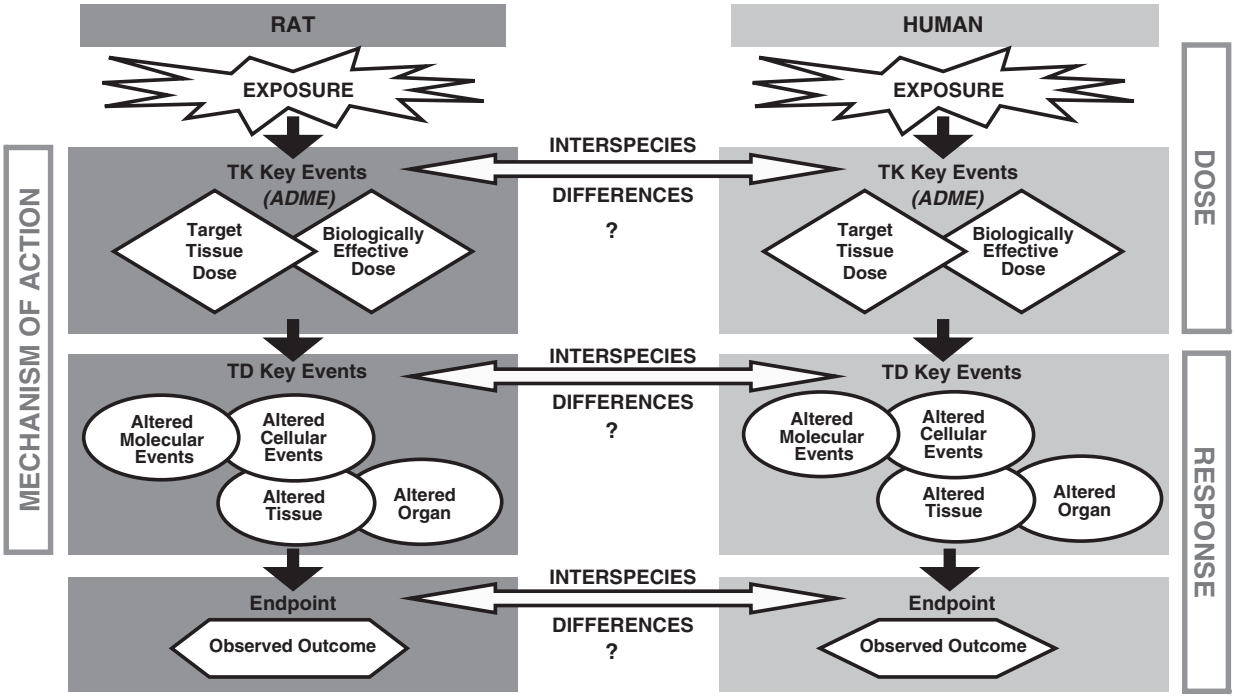

Fig. 1. Interspecies differences in chemical mechanism of action after environmental chemical exposure. Illustration of where differences between two species, rat and human, may exist in response to chemical exposure. Human and rat are shown as example species because these species are often compared in human health risk assessment; other species would be relevant examples for ecological risk assessment. Arrows with question marks represent potential biological differences in the toxicokinetic (TK) and toxicodynamic (TD) mechanisms, and in the potential health effect/outcome. (ADME = absorption, distribution, metabolism, and excretion.) There are multiple, different scenarios encountered for different chemicals in which some data on TK and/or TD are available for more than one species. For example, when comparing species for a chemical's mechanism of action where the outcome of interest is similar across species, the differences in mechanism of action could be at the TK and/or TD level.

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

inherent in extrapolating inferences across species in both human health risk assessment and ecological risk assessment.

In regards to human health risk assessment, it is usually the case for a given chemical that there are limited data available to establish human relevance (e.g., epidemiological studies, human *in vitro* assays). In the absence of human data, the assumption that health effects observed in animal test models are similar to health effects in humans is employed by EPA (for example, neurotoxicity risk assessment guidelines (U.S. EPA, 1998). Indeed, the current risk assessment paradigm often relies on data from high dose-based animal toxicological assays and on multiple extrapolations – from endpoints in one species to endpoints in another species, from high doses to low doses, and from one population to another. These extrapolated inferences about the health effects of chemicals are some of the sources of *uncertainty* inherent in risk assessment. Furthermore, like traditional toxicology studies, many of the computational toxicology models currently being developed, such as virtual tissue models, are also based on data from animal studies (www.epa.gov/ncct/virtual_liver; www.epa.gov/ncct/v-Embryo; Shah and Wambaugh, 2010). For application to human health risk assessment, inferences based on these types of models will therefore also require extrapolating across species.

To transition to the NRC and EPA vision of toxicity "pathway-based" risk assessment, the development and adoption of methods for recognizing perturbations to pathways, for identifying "toxicity pathways," and for comparing effects of toxicity pathway perturbations across species will be crucial. It is expected that the increasing availability of genome-wide data, genomics tools, and molecular mechanistic data will enable the development of such methods, including new approaches for informing interspecies differences in chemical mechanism of action (National Research Council, 2006, 2007; Benson and Di Giulio, 2007). To fully leverage the floodgate of molecular information that has opened in the past decade, environmental chemical risk assessors will need a toolbox of tested methodologies that can be used to inform interspecies differences. Currently, since environmental chemicals commonly lack a sufficient human data set, it can be difficult to assess the relevance of health effects observed in animal-based toxicology studies to human health risk assessment. Likewise, for species of concern in ecological risk assessments, approaches for better identifying differences, characterizing uncertainties, and extrapolating inferences about the toxic effects of environmental chemicals across multiple, differing species are needed.

To help advance research and development in this area of environmental chemical risk assessment, in this article we compare and contrast existing approaches that can be used, modified, and combined in new ways to discover and characterize interspecies differences in chemical mechanism of action. The focus of this article is on the critical review of the available comparative methods, and while some chemical examples are provided to highlight the use of some methods, this article does not include a comprehensive review of chemical examples. First, from disciplines as varied as evolutionary biology, systems biology, mathematics, and computer science, we compile and discuss methodologies for making simple to complex comparisons between and among sets of information. These methodologies can be applied to comparative analyses of molecules and molecular mechanisms, including comparisons across species of genomics data such as gene sequences, protein sequences, mRNA transcript abundance, protein–protein interactions, gene-protein interactions, *etc.* Second, we discuss the advantages and limitations of these various comparative approaches. Third, we assess some of the opportunities for applying these approaches within the current risk assessment structure, and we identify areas in need of further development for use in toxicity pathway-based risk assessment.

Approaches to comparing species at the molecular level

In human health risk assessment, we often need to extrapolate inferences about biological function and impairment thereof from non-human animals to humans, with the end goal of understanding the human health impacts of environmental contaminants. The mechanistic differences in responses to chemicals among different species can occur at any point along the different levels of biological organization in the continuum from exposure to endpoint (Fig. 1). To improve interspecies extrapolations in risk assessment and to reduce and characterize their associated uncertainties, we therefore need the ability to compare these differing biological levels of information across species. Thanks to genomics, bioinformatics, and computational algorithms from various fields, this ability exists. The data from genome projects and the tools of molecular biology have greatly expanded our discovery and understanding of physiological traits on a cellular and molecular basis, also enabling the examination of chemical perturbations on these differing biological levels of organization (Mattes, 2006). And, the conceptual framework of evolutionary biology plus mathematical algorithms from graph theory and complex network theory have provided the theoretical underpinnings and computational mechanisms needed for making comparisons and inferences across species. Here, we provide a compilation and overview of existing approaches that can be utilized for interspecies comparisons of genomic information. These approaches were developed within varied disciplines for varied purposes but are binned here by their relevance to differing levels of biological complexity — from genes and proteins to pathways to complex biological networks.

Gene/protein level approaches

Scientific comparisons across species in biology started with attempts to classify organisms based on their overall similarity of observable traits, typically morphological traits. This comparison of shared physical characteristics across species, known as phenetics, has largely been superseded by cladistics (Hennig, 1965), in which shared derived characters are compared to infer the evolutionary relationships among organisms. These relationships are depicted by cladograms, or evolutionary trees. As it became possible to reconstruct evolutionary relationships among species using molecules as characters, cladistic analysis using comparisons of DNA, RNA, and protein sequences (molecular phylogenetics) became the norm for testing taxonomic hypotheses. Therefore, the earliest and simplest molecular methods of inter-species comparisons were developed at the level of proteins, genes, and non-coding DNA sequences.

The computationally intensive nature of aligning, analyzing, and reconstructing evolutionary trees from molecular sequence data of multiple species gave rise to various computational sequence alignment and phylogenetic analysis methods. Because the processes of identifying the optimal multiple sequence alignment and/or the optimal tree can be prohibitively computationally expensive, these methods typically employ heuristic algorithms and scoring functions. A plethora of computer programs have been developed for both multiple sequence alignment and for phylogenetic tree reconstruction; for a fairly comprehensive listing of available programs and servers, see evolution.genetics.washington.edu/phylip/software.html. Some phylogenetic analysis methods, such as maximum parsimony, implicitly invoke an evolutionary model; some, such as maximum likelihood and Bayesian inference, explicitly employ models of evolution; and others, such as neighbor-joining, are based on genetic distance measures typically calculated using population genetics models of genetic mutation versus genetic drift, or percent sequence similarity (for reviews of phylogenetic methods, see Nielsen, 2005; and also see Yang, 2006). In addition, further advancing our abilities to make comparisons among species at the molecular level are the more recent technological breakthroughs that have given rise to high-throughput genome sequencing, and the availability of genomics data through open access online databases (see Table 1).

In the simplest scenario, the biological unit of interest for interspecies comparisons at the molecular level is a DNA sequence such as a

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

gene, or a protein sequence (see Frazer et al., 2003 for a review of methods). For example, given a gene or protein known to be involved in the mechanism of action of a toxicant in rats, we could compare its sequence to the sequence of its orthologous gene/protein in humans. Orthologs are genes/proteins in different species that originated from a common ancestral gene/protein separated by a speciation event, as opposed to genes or proteins that are derived from gene duplication events within a species. The usefulness of an ortholog comparison approach for risk assessment lays in an assumption that evolutionary conservation of a gene or protein sequence correlates with conservation of its function, which in turn corresponds to a conserved mechanism of action. Ortholog comparisons have proven useful in drug target prediction in the field of pharmacology (Searls, 2003; Gunnarsson et al., 2008). However, while evolutionarily conserved sequences are likely functionally conserved as well, recent analyses of genomic data have revealed that many sequences serving a conserved function are not themselves conserved (Margulies et al., 2007; Monroe, 2009; Wang and Zhang, 2009). Therefore, while there is support for the assumption that highly evolutionarily conserved DNA and protein sequences reveal a selective constraint that belies an important and preserved function, the absence of conserved sequence across species does not guarantee the absence of a conserved function. In other words, while conserved sequences and high non-synonymous to synonymous substitution rate ratios in proteins are good indicators of natural selection that is maintaining a function by 'constraining' sequences from diverging over time, natural selection and the functional constraints it imposes may not always be evident by examination of gene and protein primary sequence information.

Since our goal is to make inferences across species about the function and/or relative expression of a gene or protein and, thereby, its role in the mechanism of action of a toxic agent, indices of overall sequence similarity, such as percent sequence similarity or percent identity (PID) (Raghava and Barton, 2006), are of limited practical use as stand alone methods. Put another way, although gene and protein sequence similarity between species are quantifiable, such measures do not provide insight into differences in the function of those genes and proteins and how they may or may not affect the mechanism of action. Targeted approaches that compare changes within functional domains, protein structure, or gene regulatory sites, in addition to methods that compare changes in levels of protein expressed, are more likely to be useful in extrapolating inferences about gene and protein function across species. Of course, in cases of genes and proteins with known roles in xenobiotic metabolism or other well-studied disease processes, interspecies sequence differences among orthologs may already be documented and understood (see Mattes, 2006 for a review), and thus could be used in risk assessment. For example, the pesticide vinclozolin is known to act as an anti-androgen, binding to the androgen receptor (AR), thereby inhibiting androgen action by blocking binding of AR with endogenous androgens. Since this and other chemicals have been found to have such specific effects on one protein, approaches that compared the AR sequence and its binding affinity across species have been fruitful (Hartig et al., 2007; Wilson et al., 2007). In the absence of such information, several *in silico* methods for determining the structure and assessing the functional importance and functional similarities of proteins have been developed.

Because conserved sequences are good indicators of natural selection, several methods of comparative sequence analysis have been developed to find these conserved sequences across genomes and, by inference, identify functionally important regions. Regions of DNA identified this way include both protein-coding and non-

Table 1
Selection of widely-used databases useful for comparative toxicogenomics.

Publicly available database	Description	Website
BioCarta	Pathway database with graphical models	http://www.biocarta.com
Biomolecular Object Network Databank plus (BONDplus)	Integrates public and proprietary gene sequences and interactions (includes GENESEQ and BINDplus)	http://bond.unleashedinformatics.com
Comparative Toxicogenomics Database (CTD)	Manually curated data describing interspecies chemical–gene/protein interactions and chemical– and gene–disease relationships	http://ctd.mdibl.org
Conserved Domain Database (CDD)	Multiple sequence alignments of conserved protein domains drawing from several databases	http://www.ncbi.nlm.nih.gov/cdd/cdd.shtml
Database of Interacting Proteins (DIP)	Manually curated protein–protein interactions	http://dip.doe-mbi.ucla.edu
Ensembl	Genome browser for sequenced genomes	http://www.ensembl.org
Expert Protein Analysis System (ExPasy) Proteomics Server	Protein sequences and structures. (includes other databases such as UniProtKB, PROSITE, ENZYME, HAMAP)	http://www.expasy.org
GenBank	Annotated collection of publicly available sequences	http://www.ncbi.nlm.nih.gov/genbank
Gene Ontology (GO) database	Standardized representation and annotation of gene and gene product attributes using controlled vocabulary	http://www.geneontology.org
Gene Expression Omnibus (GEO)	Functional genomics data of array- and sequence-based data	http://www.ncbi.nlm.nih.gov/geo
Homologene	Automated detection of homologs among several completely sequenced eukaryotic genomes	http://www.ncbi.nlm.nih.gov/homologene
IntAct	Database system and analysis tools for protein interaction data	http://www.ebi.ac.uk/intact
IntNetDB	Computationally predicted database of human protein–protein interactions (PPIs) and their worm, fruitfly and mouse interologs	http://hanlab.genetics.ac.cn/IntNetDB.htm
KEGG PATHWAY	Collection of manually drawn pathway maps representing molecular interaction and reaction networks	http://www.genome.jp/kegg/pathway.html
KEGG DISEASE	Collection of disease entries capturing knowledge on genetic and environmental perturbations	http://www.genome.jp/kegg/disease
Nature Pathway Interaction Database (PID)	Database of molecular interactions and biological processes assembled into pathways	http://pid.nci.nih.gov
Online Inheritance in Man (OMIM)	Compendium of human genes and genetic phenotypes	http://www.ncbi.nlm.nih.gov/omim
Reactome	Manually curated and peer-reveiwed database of pathways that includes a species comparison tool	http://www.reactome.org
RCSB Protein Data Bank (PDB)	Archive of information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies	http://www.pdb.org
SCOP database	Comprehensive database of the structural and evolutionary relationships among all proteins whose structure is known	http://scop.mrc-lmb.cam.ac.uk/scop
STRING	Known and predicted protein-protein interactions	http://string.embl.de/
ToxCast	High throughput assay data for chemical compounds	http://www.epa.gov/comptox/toxcast
UCSD genome browser	Genome browser for sequenced genomes	http://genome.ucsc.edu

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

coding elements. The comparative genomics methods used to find these evolutionarily conserved regions tend to incorporate and build on one or more phylogenetic analysis methods. For example, one type of approach uses maximum likelihood gene or protein trees to infer relative evolutionary rates among multi-species sequence alignments (e.g., the "Evolution–Structure–Function" (ESF) analyses of Simon et al., 2002); another type develops statistically rigorous single-site measures of evolutionary conservation using maximum likelihood analysis and models of nucleotide substitution (e.g., the "Genomic Evolutionary Rate Profiling" (GERP) of Cooper et al., 2005; Goode et al., 2010); and yet another type makes use of hidden Markov models whose parameters are estimated from multiple alignment by maximum likelihood. Finally, by integrating the results of several such alignment and conservation-detecting approaches, Margulies et al. (2007) showed that robust annotations of selectively constrained genomic regions could be produced.

As already mentioned, and as found to be true for the toxicologically important nuclear receptor PXR (pregnane X receptor) (LeCluyse, 2001; Moore et al., 2002; Mattes, 2006), some proteins maintain a conserved function across evolutionary time without maintaining highly conserved primary sequences. To find these types of proteins, Barthet and Hilu (2008) developed an approach that identifies putative functional domains within rapidly evolving protein sequences. Their method evaluates the constraint on amino acid side chain composition across the open reading frames among multiple species. The basic assumption of this approach is that functional conservation across species is indicative of selective constraints on function; and, if selection is not acting on and therefore not evident in the primary DNA or amino acid sequence of a functionally conserved protein, then it must be acting on another level, such as on protein structure or on the physicochemical properties of particular amino acids. Additional methods for comparing the physicochemical variation at individual amino acid sites in proteins had previously been developed in the context of predicting the functional impairment or difference in function of proteins. Such methods either qualitatively examine (e.g., SIFT — Sorting Intolerant From Tolerant; Ng and Henikoff, 2003; Kumar et al., 2009) or quantify (e.g., MAPP — Multivariate Analysis of Protein Polymorphism; Stone and Sidow, 2005) the physicochemical variation and functional impact of amino acid differences at each site in orthologous proteins using multiple species sequence alignments and, in the case of MAPP, evolutionary trees inferred from them to control for phylogenetic correlation.

Finally, in addition to comparing primary sequences and the physicochemical properties of protein residues, it is possible to directly compare the three-dimensional (3D) structures of proteins to infer differences in function across species. A number of algorithms that automate protein structural alignment and comparison using a variety of criteria have been developed (e.g., Taylor and Orengo, 1989; Falicov and Cohen, 1996; Shindyalov and Bourne, 1998; Jewett et al., 2003). Examples of online resources that implement such algorithms, find and predict protein structures, and make 3D comparisons include: the Structural Classification of Proteins (SCOP) database (scop.mrc-lmb.com.ac.uk/scop) (Murzin et al., 1995; Lo Conte et al., 2000; Andreeva et al., 2004; Andreeva et al., 2008) with its associated collection of manually curated structural alignments, SISYPHUS (Andreeva et al., 2007); the ASTRAL Compendium for Sequence and Structure Analysis (astral.berkeley.edu) (Brenner et al., 2000; Chandonia et al., 2002, 2004); MATRAS Protein 3D Structure Comparison (biunit.naist.jp/matras/) (Kawabata, 2003); and, MinRMS: A Tool for Determining Protein Similarity (www.cgl.ucsf.edu/research/minrms) (Huang et al., 2000; Jewett et al., 2003). Additional tools and resources can be found on the RCSB Protein Database (RCSB PDB) archive (www.pdb.org) (Berman et al., 2000). However, most of these protein structure comparison methods were created at a time when structural analysis was the primary means of determining homologies between remotely related proteins. Due to the explosion of genomics technologies and sequence databases, protein homologies are now most efficiently determined based on sequence analysis. Indeed, even the SCOP database now relies on integrating sequence-based information into their protein comparison and classification scheme (Andreeva et al., 2008). Additionally, there is some evidence that the functional domain architecture of proteins is more likely the result of evolutionary descent, or phylogenetic legacy, than selection for function (Gough, 2005). This implies that methods incorporating the evolutionary relationships among proteins being compared will fare better in the context of understanding functional similarities and differences than comparisons based on structure alone. Therefore, 3D protein comparison across species is not currently a very useful approach for informing interspecies similarities and differences in protein function.

Chemical examples. To predict human effects, a number of studies have performed genomic assessments of the interspecies differences in liver toxicity, a common side effect of certain pharmaceuticals and chemicals. One chemical example is the pharmaceutical, acetaminophen, which can be hepatotoxic and can cause death when used at overdose. The mechanism of action for the hepatotoxic response has been investigated in genomic studies in a number of species including human cells (reviewed in Mattes, 2006). The hepatotoxicity is thought to be a result of exposure to the acetaminophen metabolite, N-acetyl-*p*-aminophenol (APAP). To identify a hepatotoxic genomic signature, Beyer et al. (2007) conducted a multicenter investigation comparing genomic changes induced in rat and mice livers (*in vivo*) by the hepatotoxic APAP and by a non-hepatotoxic APAP isomer (N-acetyl-*m*-aminophenol). They found that the hepatotoxic genomic signature includes *c-Myc* induction and further, developed a species-independent (i.e., in common to rat and mice) genomic signature relevant to liver necrosis that includes effects on *c-Myc* and *Jun* oncogenes. Their findings suggest some similarities in the mechanism of action between rat and mice but fails to discuss possible differences in the mechanism of action since the study did not present the rat and mouse differences in gene expression. Another chemical with hepatotoxic side effects, coumarin, was studied using an approach called an "informational bridge" to connect gene expression responses in *in vitro* rat hepatocytes, *in vitro* human hepatocytes, and *in vivo* rat livers (Uehara et al., 2010). First, Uehara et al. (2010) identified "*in vivo–in vitro* bridging probes" from the rat data. Then these probes were assigned to their human orthologous genes to identify "rat–human bridging probes" which were used to study coumarin-induced gene expression changes in human hepatocytes. The pattern of changes in 25 genes was found to be similar between rat and human cells but the degree of change was greater in the rat.

Another type of comparative genomics study that focuses on interspecies comparisons of sequence-based functional elements was utilized to identify and compare human, mouse, and rat dioxin response elements (DREs) (Sun et al., 2004). 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD; dioxin) mediates its effects via the aryl hydrocarbon receptor (AhR); dioxin binds to the AhR and alters the regulation of AhR-mediated genes. In this study, DRE-containing genes were identified and compared across species to identify common genes. The analysis found that DRE-regulated genes differed more between rodents and humans than between the two rodents. These studies provide examples of real-world interspecies comparative strategies, employed in pharmaceutical development and environmental toxicology, to predict human effects.

Pathway level approaches

Both adaptive (functional) and non-adaptive (neutral) evolutionary divergence can cause differences among species, but such differences in single genes are not typically representative of the

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

complexity of a trait, since very few traits are monogenic. In other words, in most cases, multiple genes and regulatory elements interact with each other and with environmental factors to produce a phenotype. Therefore, using interspecies comparisons of a single gene or protein as a proxy for a phenotype of interest may be an overly simplistic method for extrapolating information from one species to another. It has therefore been postulated that molecular pathway comparisons will more accurately reflect similarities and differences across species in complex phenotypes such as those involved in toxicity endpoints (National Research Council, 2009). A molecular pathway can be generally defined as a series of linked biochemical steps involving genes, proteins, and other biochemical factors whose activity results in the ebb and flow of molecules that lead to a biological effect. Any type of pathway, such as a metabolic pathway like the fatty acid synthesis pathway or a signaling pathway like the insulin receptor pathway, that is involved in the mechanism of action of a xenobiotic and can potentially lead to an adverse outcome, can be considered a "toxicity pathway". Therefore, comparisons of these pathways may be more informative biological levels for understanding the implications of interspecies extrapolation than single genes or proteins.

The most basic approach to pathway comparison is based on the presence or absence of orthologous proteins across species without regards to the direct interactions among the proteins. This approach was developed to predict the function of uncharacterized proteins, based on the assumption that proteins that co-occur across organisms likely function together, and that these clusters of proteins correspond to biologically meaningful pathways. Original presence/absence methods conduct a cluster analysis of organism-by-protein matrices that are either based on simple presence/absence binary data (i.e., 1 = protein present in that organism; 0 = protein absent) (Pellegrini et al., 1999; Liao and Noble, 2002; Wu et al., 2003) or that are modified by a measure of sequence similarity represented as continuous data (Date and Marcotte, 2003; Jothi et al., 2007). Improvements to the basic presence/absence methods extend those approaches from comparisons of co-occurring proteins to comparisons of co-evolving proteins. By incorporating the evolutionary relationships among the species being compared into their algorithms (Barker and Pagel, 2005; Zhou et al., 2006; Barker et al., 2007; Cokus et al., 2007), these methods provide additional information on the gain or loss of proteins in an evolutionary context. A related approach, also developed for the characterization of proteins and discovery of pathways, is based on comparing the size of gene families across species. As with the presence/absence approaches, some of these gene family size methods are correlation-based (Ranea et al., 2007) and others are coevolution-based (Cordero et al., 2008; Ruano-Rubio et al., 2009). Finally, there are analogous approaches that use binary presence/absence data, but that tally the presence/absence of other factors instead of tallying the pathway proteins. For example, Tun et al. (2006) analyze the presence/absence of interactions between the proteins, while Kastenmuller et al. (2009) and Gonzalez and Zimmer (2008) developed algorithms that compare pathways based on the presence/absence of phenotypes, including continuous, or quantitative, phenotypes. The latter two methods were developed to understand the biochemical basis of phenotypic traits, and they work by linking phenotypes to associated metabolic processes. Because toxicity endpoints are phenotypes, this method has potential for discovering and informing differences in toxicity pathways across species.

Unlike most of the aforementioned approaches, other types of interspecies pathway comparison take into account the interactions, or reactions, among the enzymes and compounds in the pathways being compared. One such approach analyzes the interacting functional domain patterns in protein–protein interactions across multiple species (Chen et al., 2008) — domains being the structural and functional units of proteins. The assumption underlying this method is that shared interacting domain patterns across species are indicative of shared functions; and according to the authors, this method achieves better than 95% accuracy at predicting gene function in humans. However, most of the approaches incorporating interaction information into interspecies pathway comparisons were not developed not as a means to predict and characterize the functionality of proteins and pathways. Most were created as an additional means to uncovering the evolutionary relationships among organisms, and especially among microbes whose evolutionary relationships can be intractable to standard sequence-based phylogenetic analyses due to the complicating factor of horizontal gene transfer. Approaches were therefore developed that attempt to reconstruct robust phylogenetic trees using functional pathway information (Heymans and Singh, 2003). These approaches typically model their molecular pathways as graphs (see Deville et al., 2003 for overview of graph models), align the graphs across one or more species, and then assess and score the similarity among the graphs in one way or another. For example, one web server (www.jaist.ac.jp/~clemente/cgi-bin/phylo.pl) (Clemente et al., 2007) graphs comparisons of pathway structure using different measures of protein functional similarity: measures based on enzyme class (hierarchical similarity) (Tohsato et al., 2000; Chen and Hofestadt, 2004), information content similarity (Tohsato et al., 2000; Pinter et al., 2005), or gene ontology similarity (Clemente et al., 2005). It pulls pathways from the KEGG database (www.genome.jp/kegg/pathway.html) and currently allows for comparison across 13 species including mouse, rat, and human. Other approaches assume that functional similarities are inherent in primary sequences and so use a measure of sequence similarity from either pairwise (Ovacik and Androulakis, this issue) or multi-species sequence alignments (Forst and Schulten, 1999, 2001) to compare topologically aligned pathway graphs. Even though many of these approaches were developed for studying evolution in microbes, their potential usefulness to risk assessment is in providing the capacity to align and compare pathways across species, regardless of the taxa.

Methods that highlight the differences between species in their use of similar pathways and similar reactions within pathways may be especially useful for informing interspecies differences in toxicity pathways. One such method, the Metabolic Pathway Alignment and Scoring (M-PAS) method, is another graph alignment approach but is based on discovering conserved pathways from "building blocks" of aligned reactions that account for reaction direction (Li et al., 2008). This method integrates enzyme function and sequence similarity scores into one similarity score and aligns networks of these building blocks, resulting in aligned pathways with quantified levels of conservation, giving insight into the differences between species. Although currently limited to two-species comparisons and to linear (acyclic) pathways, Li et al. (2008) suggest the M-PAS can be expanded to more complex graph topologies and to multi-species comparisons.

Chemical example. In mice, PPARα regulates hepatic lipid metabolism, but there is an absence of data on the mechanism of PPARα in humans. To address data gap, Rakhshandehroo et al. (2009) compared gene level and pathway level approaches by assessing the genomic response in primary hepatocytes exposed to the PPARα agonist pharmaceutical, Wy14643. Their comparison of mouse and human hepatocyte *(in vitro)* responses found little overlap at the gene level but a greater overlap at the pathway level. Common pathways included lipid metabolism and known PPARα target pathways (e.g., HMGCS2). In addition, human specific PPARα-regulated pathways for xenobiotic metabolism and apolipoprotein synthesis were identified. Their results suggest some common and some divergent liver responses to PPARα between rat and human. Identification of such human-specific and in-common pathway effects can improve our knowledge about species-specific mechanistic steps and aid the prediction of human health effects.

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

Network level approaches

Even pathways are relatively simplistic models of extremely complex biological processes. In reality, pathways overlap, interact with each other, and are components of highly complex non-linear molecular networks (see Wuchty et al., 2006 for a review of biological network architecture). The nodes of molecular networks typically represent molecular entities like genes, proteins, mRNA levels, or metabolites, and the edges connecting these nodes represent the functional relationships among those entities, such as protein–protein interactions, gene modifications, transcriptional regulation, etc. (Han, 2008; Schadt, 2009). Complex traits, including common human diseases, originate from more than single gene disruptions; they stem from a complex interaction between an organism's molecular networks (genetic and epigenetic) and a broad array of environmental factors including exposure to environmental toxic agents. Recent research in systems biology has shown that molecular networks both respond to and, in effect, cause common human disease through perturbations to complex interconnected molecular interactions; and that therefore, they may be more relevant to understanding diseases than are a small number of genes or single pathways (Chen et al., 2008; Schadt, 2009). Because of this, comparisons of entire molecular networks and/or of interacting sub-networks underlying or associated with toxicological endpoints should provide yet another method for informing interspecies differences in a risk assessment context.

Approaches to interspecies network comparison share similarities with methods used for pathway comparison, since both involve graphical models. However, additional challenges not encountered by most pathway comparisons include the non-linear structure of networks, the exponential growth of the size of the alignment graphs with increasing number of aligned networks, and the dynamic nature of networks (Han, 2008). In addition, molecular networks are not static; they have "dynamic modularity" in that they can vary over time and space (Kholodenko, 2006; Han, 2008). Nevertheless, several approaches to network comparison have been developed recently and, as occurred within the now mature field of sequence alignment, improvements in this area are expected to continue progressing quickly. In regards to interspecies network differences, there are two general types of relevant network comparisons: network alignment approaches, including global and local network alignment as well as network querying (see review by Sharan and Ideker, 2006), and clustering approaches.

Global and local network alignment involve the comparison of two or more networks and have mainly been used to elucidate network structure and to identify sub-networks conserved across species and therefore likely to represent functional biological units. Network querying, on the other hand, is analogous to searching a sequence database for a match to an input sequence; it involves graph-mining techniques that search for sub-networks similar to the input query in a given network or network database. These methods could be useful for informing interspecies differences in the context of chemical risk assessment both by enabling the discovery of homologous network structures and by aiding in the identification of orthologous proteins, orthologous protein–protein interaction pairs (interologs), and signaling and regulatory pathways across species.

Network alignment and querying algorithms consist of scoring functions and search algorithms. Most network alignment approaches developed to date conduct global network comparisons by scoring the homologous node pairs and their interactions, constructing network alignment graphs from these, and then searching these sets of graphs for the highest scoring alignments (e.g., Koyuturk et al., 2004; Sharan et al., 2005a, 2005b; Flannick et al., 2006; Pinney et al., 2007). These two essential parts of network alignment – a scoring framework that represents current knowledge pertaining to the network's entities and evolution, plus an algorithm for the rapid identification of conserved alignments from an exponentially large set of possible alignments – are problems with multiple solutions. For example, one solution to the scoring problem uses a probabilistic function (e.g., PATHBLAST: Kelley et al., 2003), while another employs an overall confidence score based on the combined statistical significance of sequence similarity, number of conserved interactions, and functional relatedness (e.g., HopeMap: Tian and Samatova, 2009), and yet others adopt evolution-based scoring (e.g., MaWISh: Koyuturk et al., 2006; and, Graemlin: Flannick et al., 2009) or incorporate probabilistic models of evolution (e.g., Berg et al., 2004; Berg and Lässig, 2006; and, CAPPI: Dutkowski and Tiuryn, 2007, 2009). Continued improvements to scoring functions, including more phylogeny-based approaches, are expected as our understanding of molecular network evolution increases.

As with scoring functions, several approaches to the problem of the search algorithm have been developed. However, while the scoring problem is in the realm of statistics, probability, and evolutionary and molecular biology, the search algorithm problem sits squarely in the fields of graph theory and in the latest extension of graph theory: complex network theory. Solutions to the search problem have to address the issues of speed, scalability, and accuracy. Most current solutions to the search problem include approaches, such as progressive alignment, that use a stepwise, iterative algorithm for finding locally optimal solutions based on the score (a "greedy heuristic"). A heuristic is employed because the growth of the alignment graphs is a limiting factor in terms of computational cost. Indeed, one large hurdle for many of these approaches has been in scaling up beyond pairwise comparisons (Kelley et al., 2003; Koyuturk et al., 2004, 2006; Sharan et al., 2005a; Zaslavskiy et al., 2009) to multi-species network comparisons (Sharan et al., 2005b; Flannick et al., 2006, 2009; Alexeyenko and Sonnhammer, 2009; Dutkowski and Tiuryn, 2009; Qian and Yoon, 2009; Tan et al., 2009). In addition, a non-heuristic pairwise comparison solution to the search algorithm problem has also recently been proposed by Klau (2009). According to the author, this algorithm is adaptable to many types of networks and is scalable to multi-network comparisons; it works by mathematically redefining the global alignment graph in such a way that a search for the optimal structural alignment is possible.

Creative modifications of these approaches, as well as alternatives to network alignment approaches, have also recently been developed. For example, whereas most alignment graphs are composite graphs in which nodes are merged representations of orthologous relationships among molecules like genes and proteins connected by edges representing interactions, there are network alignment approaches that instead define the nodes as the interactions or as sets of interactions. One such approach, the "direct-edge-alignment" method, uses domain–domain interactions across two networks to construct alignable pairs of edges and thereby directly infers a protein–protein interaction network (DOMAIN: Guo and Hartemink, 2009). Another approach creates networks of interacting pathways where the nodes are pathways and the edges between them are the compounds they exchange or share (Mazurie et al., 2008). On the other hand, network comparison approaches that do not even use network alignment to address interspecies questions have also been developed. Because network alignment can be limited in coverage if dependent on orthologous relationships, and because it is potentially sensitive to errors in network topology, some investigators have instead relied on graph clustering algorithms for comparisons across networks. For example, there are connectivity-based clustering approaches (e.g., Bergmann et al., 2003; Yamada et al., 2006; Narayanan and Karp, 2007; Borenstein et al., 2008; Erten et al., 2009; Wiles et al., 2010), data integration clustering approaches (Kelley and Ideker, 2005; Linghu et al., 2008; Alexeyenko and Sonnhammer, 2009; Wang et al., 2009), and probabilistic clustering approaches (Stuart et al., 2003) that have successfully combined and compared a variety of data types across multiple species. In addition, at least one approach integrates network alignment methods with graph clustering algorithms to compare protein functional similarity across species (Ali and Deane, 2009).

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

Discussion

The tools and concepts of comparative genomics, including genome-wide measurements of biological entities such as proteins, expressed genes, metabolites, and DNA methylation (i.e., proteomics, transcriptomics, metabolomics, and genomic methylation), have the power to greatly enhance toxicology and risk assessment. Comparative genomics can provide a means for identifying toxicity pathways, mechanisms, and biomarkers, and their differences across species. In the toxicity pathway-based approach to risk assessment, with its envisioned transition to a greater reliance on in vitro assays, comparative genomics will be crucial for establishing mechanistic links and quantitative relationships between laboratory animal studies and human cell-based in vitro assays. Thus, the comparative genomics and computational approaches compiled here will be critical to addressing one of the challenges in human health risk assessment: reducing the uncertainty caused by extrapolated inferences about toxicological exposure and health effects from test animals to humans (Fig. 1).

It is beyond the current state of the science to produce a guide on using genomics for extrapolating inferences across species in environmental chemical risk assessment. Instead, by categorizing the different types of approaches, starting with interspecies comparisons at the relatively simplistic level of genes and proteins and ending with comparisons at the systems biology level of molecular networks, we have presented an overview of the main types of approaches available to researchers (Fig. 2). We also provide examples of several widely used genomics databases and publicly available interspecies comparison tools and resources in Tables 1 and 2 respectively. In this section, we evaluate the strengths and weaknesses of these major types of interspecies molecular comparison approaches in terms of their potential for implementation in risk assessment. Furthermore, we discuss the challenges and overarching recommendations for research that will help bridge the gap between current comparative genomics approaches and application to risk assessment.

Strengths and weaknesses of approaches

Even though the quantity of available molecular information for different species is increasing rapidly, we face a variety of technical and experimental issues when using these data in interspecies comparisons. Genome-wide data are full of noise (false positives) and annotation from one species to another can be unreliable. Annotation of genomes and experimental arrays, as well as the availability of genomic data and experimental platforms, are technical issues that impact quantity, quality, and completeness of data sets. With pathway and network comparisons in particular, we have to deal with this dilemma of incomplete data sets when evaluating whether or not the comparisons accurately represent the systems being modeled. For example, given a protein–protein interaction network, if the set of interactions being examined is incomplete, then the absence of an observation assumed to indicate non-interaction is potentially erroneous. In addition to incomplete data sets, we also have the problem of inconsistent experimental conditions across studies. Differences across studies in coverage of genome-wide assays, in experimental design, and in analyses can impact conclusions drawn when comparing results from such studies. Therefore, comparison methods that are robust to data gaps and inconsistencies such as these are needed. To account for shortcomings in a way that is useful in the context of risk assessment, performance estimates of the models and comparison approaches used will be critical.

Of all levels of interspecies molecular comparison, the comparison of single entities such as DNA sequences, protein functional domains, whole protein structures, or even single amino acid mutations, has the longest history and therefore is the best-tested set of methodologies. Sequence comparison methods are robust, well-understood, and

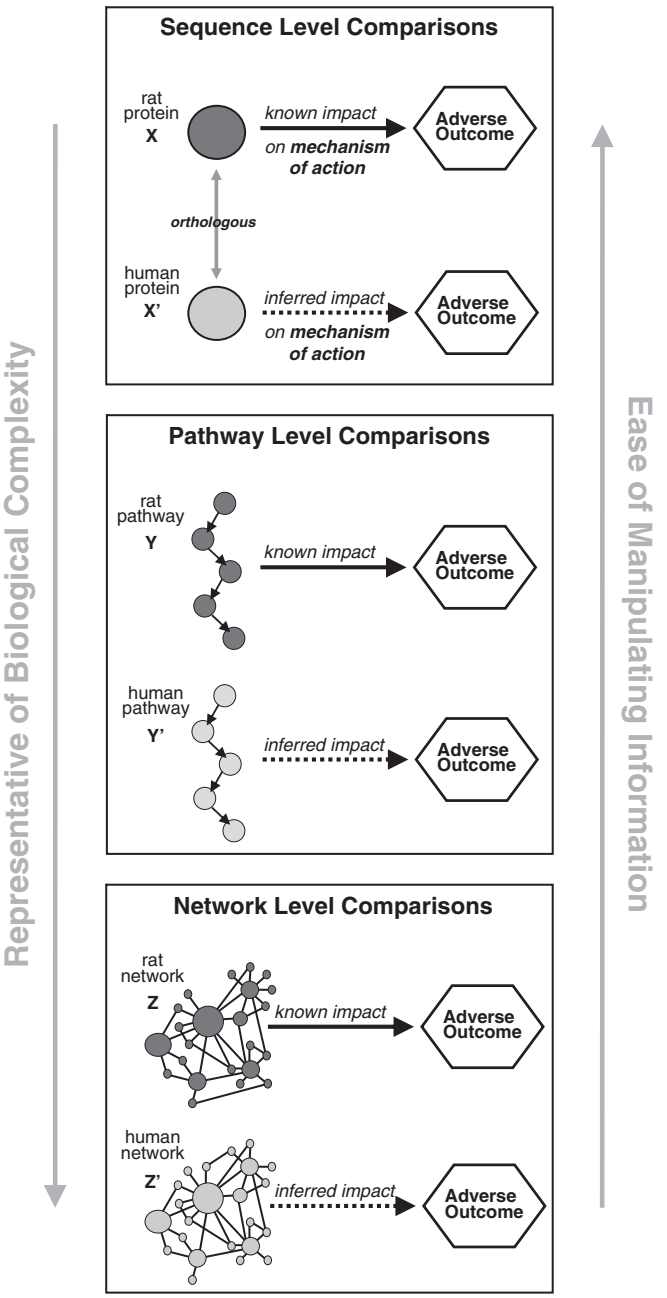

Fig. 2. Molecular-based interspecies comparisons. Simplified depiction of three levels of molecular comparisons for extrapolating inferences about mechanism of action across species (showing rat and human as examples) using comparative genomics. In reality, pathway models tend to involve more nodes and interactions than depicted, and may also have non-linear components; in addition, molecular network models using genome-wide data will be far more complex in terms of nodes and interactions/linkages than depicted in the cartoon above, giving them the appearance of "hairballs".

widely used. While minor issues with ascertaining orthology do pertain to these approaches, they do not suffer from the larger problem of incomplete datasets. In addition, techniques to predict functional changes based on sequence or structural differences have proven useful in hypothesis generation and weight-of-evidence schema in the context of gene characterization and discovery. However, it is not easy to predict how a functional change in a sequence, such as an enzyme critical to the mechanism of action, may impact a toxicological outcome. Although the reductionist approach is comfortable and attractive in both science and risk assessment, the isolation of complex biological problems, like toxicity, to the level of a gene or protein may mean missing the mark more often than not. In other words,

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

Table 2
Selection of publicly available online tools for interspecies molecular comparisons.

Tool/resource	Description
Protein & gene comparison approaches	
GERP	*Genomic Evolutionary Rate Profiling*: identifies constrained elements in multiple sequence alignments by quantifying substitution deficits (Cooper et al., 2005)
http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	
MAPP	*Multivariate Analysis Of Protein Polymorphism*: predicts constraint and the impact of polymorphisms; maps these onto predicted protein structures (Stone and Sidow, 2005)
http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	
MATRAS	*MArkovian TRAnsition of Structure evolution*: program for 3-D protein structure comparison (Kawabata and Nishikawa, 2000; Kawabata, 2003)
http://biunit.aist-nara.ac.jp/matras	
MINRMS	*MINimal Root-Mean-Squared-Distance*: Tool for determining protein similarity through 3D alignment (Jewett et al., 2003)
http://www.cgl.ucsf.edu/Research/minrms	
PANTHER	*Protein Analysis THrough Evolutionary Relationships* (Thomas et al., 2003; Mi et al., 2005)
http://www.pantherdb.org	
PHAST	*PHylogenetic Analysis with Space-Time models*: Freeware package for comparative and evolutionary genomics (Siepel et al., 2005)
http://compgen.bscb.cornell.edu/phast	
ProPhylER	*Protein Phylogeny and Evolutionary Rates*: Predicts the impact of coding polymorphisms using phylogenetic conservation (Binkley et al., 2010). Includes MAPP algorithm. (Stone and Sidow, 2005)
http://www.prophyler.org	
RIO	*Resampled Inference of Orthologs*: Estimates the reliability of protein orthology using bootstrap values (Zmasek and Eddy, 2002)
http://rio.janelia.org/	
SIFT	Predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids (Ng and Henikoff, 2003; Kumar et al., 2009).
http://sift.jcvi.org	
SISYPHUS	Structural 3D alignments for proteins with topological irregularities (Andreeva et al., 2007)
http://sisyphus.mrc-cpe.cam.ac.uk/sisyphus	
Pathway Comparison Approaches	
GeneCensus	Database tool for comparing genomes, including pathway analysis of activity levels. Includes Tree Viewer & PathwayPainter modules (Lin et al., 2002)
http://bioinfo.mbb.yale.edu/genome	
PATHBLAST	Global alignment graph method (Kelley et al., 2003)
http://www.pathblast.org	
Phylogenomic reconstruction from non-genomic data	Graph comparisons of metabolic pathway structure using different measures of protein functional similarity (Clemente et al., 2007)
http://www.jaist.ac.jp/~clemente/cgi-bin/phylo.pl	
Network comparison approaches	
Cytoscape	Open source bioinformatics platform used for analyzing and visualizing complex graph data. Free and user-friendly network visualization tool (Sharan et al., 2007)
http://www.cytoscape.org	
DOMAIN	Pairwise network alignment for protein-protein interaction networks that uses a direct-edge alignment paradigm based on domain–domain interactions (Guo and Hartemink, 2009)
http://www.cs.duke.edu/~amink/software/	
Graemlin	Network alignment framework for finding conserved modules in a set of networks or for finding matches to a particular module within a databse of interaction networks; 2.0 version includes a parameter learning algorithm for the scoring function. (Flannick et al., 2006; Flannick et al., 2009)
http://graemlin.stanford.edu	
IsoRank & IsoRankN	Tools for global multiple network alignment of protein–protein interactions for yeast, fly, worm, mouse, and human. IsoRank algorithm is similar to Google's PageRank (Singh et al., 2008); IsoRankN tool is based on spectral clustering, is error-tolerant, and computationally efficient (Liao et al., 2009)
http://groups.csail.mit.edu/cb/mna/	
NeMo	A Cytoscape Plug-in that identifies network modules (Rivera et al., 2010)
http://baderlab.bme.jhu.edu/baderlab/index.php/NeMo	
NetGrep	Fast network schema searches in interactomes (Banks et al., 2008)
http://genomics.princeton.edu/singhlab/netgrep/	
NetworkBLAST	Analyzes protein interaction networks across species to infer evolutionarily conserved protein complexes; capable of multiple network comparisons (Sharan et al., 2005a; Kalaev et al., 2008)
http://www.cs.tau.ac.il/~bnet/networkblast.htm	
PATHICULAR	A Cytoscape Plug-in that identifies regulatory path motifs, enriched paths in integrated physical networks connecting cause-effect genes in perturbational expression data (Joshi et al., 2010)
http://bioinformatics.psb.ugent.be/software/details/pathicular	

these small entities do not exist and operate in isolation; their connections and interactions with other molecules may have ramifications beyond their perceived functional change; or on the other hand, their connections and interactions could provide resiliency — a damping of effects. However, in addition to computational comparisons of genes, proteins, and other sequences from genomic datasets, there are also targeted experimental approaches that can address and help quantify interspecies comparisons at this biological level of organization, such as relative expression, gene knock-down, and enzyme activity assays (see Benson and Di Giulio, 2007 for overview). Because of this, we expect that many initial attempts to incorporate interspecies molecular comparisons into risk assessment and to quantify differences will be attempted at this level.

More biologically relevant modeling of toxicological perturbation will happen at the level of pathways and networks. These "systems" may have resiliency and may have weaknesses that are not evident in comparisons of single gene or protein differences. Pathway comparisons fit nicely into the NRC envisioned framework of toxicity pathway-based risk assessment. They provide a physiological context for protein and gene changes in a model that may better capture the key events of a mechanism of action. Approaches for comparing pathways and for measuring conservation of pathway function are still developing and are still mainly being used for understanding prokaryote evolution. While such approaches provide the structure needed for interspecies pathway comparison and are becoming better at extracting information about evolutionary relatedness based on functional systems, they do not yet encode a means for quantifying functional changes or their physiological implications.

Networks are a natural extension of pathway models; they are as unbiased a molecular representation of a biological system as is possible. In addition, because pathways are essentially simplified networks, there is no reason that network comparison approaches cannot be applied to them — there simply has been a separation of goals driving the two approaches. Networks can model anything we can measure including the functional relationships among molecules such as: gene modifications, protein–protein interactions, transcriptional regulation, DNA methylation, pre- and post-translational modifications, metabolic reactions, siRNA–mRNA interactions, etc. Because molecular network models of genome-wide processes are all-encompassing in nature and constructed from unbiased empirical

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

data, and because they display an evolved robustness to perturbations (Albert et al., 2000; Wuchty et al., 2006; Zhu et al., 2007; Costa et al., 2008), molecular networks are the most comprehensive and biologically relevant models available for understanding physiological responses to environmental chemicals and are expected to greatly enhance 21st century risk assessment (Edwards and Preston, 2008). However, among the approaches discussed here, interspecies network comparisons have the potential to suffer the most from the problem of incomplete information resulting from disparities in availability and annotation of data across, as well as within, species. Nevertheless, a low percentage of missing or mis-annotated pieces of information from a complex model of a biological system such as a molecular network, are less likely to affect overall analyses and conclusions than are the unknown and missing data left out due to the simplicity of models such as comparisons of cherry-picked genes and proteins.

Challenges and recommendations for risk assessment

The interspecies comparison approaches compiled and discussed here were all developed for purposes other than risk assessment, such as for inferring evolutionary relationships, for predicting functional sequences and conserved pathways, for hypothesis generation in finding and testing disease-related processes, *etc*. The integration of genomics-based approaches such as these with *in vivo* and *in vitro* toxicity testing data will enhance the identification of toxicity pathways, improve our understanding of similarities and differences in mechanism of action among species, and will also aid in the development of virtual tissue models. A major challenge for the future paradigm of human health risk assessment is in translating approaches like these into ways of extrapolating inferences qualitatively, and ultimately quantitatively, about toxicological outcomes from test animals to humans. Addressing this challenge will require the development of genomics-based risk assessment methods. We propose that, with further refinement and testing, computational interspecies comparison approaches, such as those discussed here, can be adapted for this use.

Additionally, molecular data can be incorporated into the current paradigm of risk assessment as well. Information on molecular differences between test animals and humans can be applied qualitatively by using a weight-of-evidence approach to judge whether or not to adjust the *interspecies uncertainty factor* (Box 1) and/or whether or not the animal data are relevant to humans. Data from gene/protein, pathway, and network comparisons could contribute to the weight of evidence about interspecies differences for a mechanism of action that may already include, for instance, known differences in the outcomes and/or regulation of the outcome across species. For example, in the case of chemicals that affect androgen action, extensive information known about the regulation of male sexual differentiation by androgens across vertebrates (Hotchkiss et al., 2002; Wilson et al., 2004, 2007; Hartig et al., 2007) would be part of the weight-of-evidence that, combined with gene, pathway, and/or network findings, may inform a picture of high or low similarity of the mechanism of action across test species and humans. In this qualitative application, the degree of change in the interspecies uncertainty factor would be based on scientific judgment and precedent. Indeed, current and future risk assessment decision-making processes are based, in part, on scientific judgment after weighing multiple factors including data sources, data coverage, data quality, and exposure scenarios.

The larger challenge is in coming up with methods to utilize the findings of these comparative approaches in a quantitative manner. For example, while interspecies differences between gene and protein sequences are easy to quantify, the functional differences and phenotypic differences they may or may not cause are *not* easy to quantify. This difficulty in quantifying functional and phenotypic divergence can then be exacerbated by a lack of information on the relative importance of the genes or proteins to the chemical mechanism of action. On the other hand, less reductionist approaches, such as those involving pathway and network comparisons across species, show more potential for eventual quantitative application to risk assessment, because they have the capacity to measure perturbations on a systems level. However, for quantitative application, methods shown to be robust to data gaps and data inconsistencies and that include performance estimates are still needed.

Even without methodologies refined specifically for application to risk assessment, interspecies network and pathway perturbation models based on existing methods can be used now to generate testable hypotheses for human cell-based *in vitro* assays that can then be used qualitatively, and eventually quantitatively, in risk assessment. For example, as illustrated in Fig. 3, given chemical X with known adverse outcomes in rat *in vivo* studies: first, a rat network perturbation model could be developed; second, the rat and human networks could be compared; third, based on differences and similarities highlighted by the interspecies network comparison, human *in vitro* assays could be developed and performed; fourth, *in vitro* perturbations could be mapped back to the compared networks; and, finally, human *in vivo* outcomes could be inferred. Furthermore, the pathway and network comparison approaches discussed here that do not depend on evolution-based scoring or phylogenetic information could also be adapted to address another challenge in the new risk assessment paradigm: extrapolating perturbation data from *in vitro* assays conducted in one cell or tissue type to the cells, tissues, and/or organs relevant to the critical effect *in vivo*. The pharmaceutical example of the study of Wy14643 (Rakhshandehroo et al., 2009)

Box 1

Uncertainty factors in human health risk assessment

Although risk assessment guidelines and methods vary, the issue of uncertainty pertains regardless. As one example of accounting for uncertainty, the EPA's Integrated Risk Information System (IRIS) applies numerical correction factors to the reference doses (RfDs) and reference concentrations (RfCs) in human health risk assessment. Different *uncertainty factors* are used to account for: 1) variation in susceptibility among humans (intraspecies uncertainty; see Mortensen and Euling (this issue) for a review of approaches); 2) uncertainty in extrapolating from effects in animal models to effects in humans (interspecies uncertainty); 3) uncertainty in extrapolating from subchronic exposure data to chronic exposure; 4) uncertainty in making inferences based on a lowest-observed-adverse-effect-level (LOAEL) instead of a no-observed-adverse-effect-level (NOAEL); and, 5) uncertainty due to incomplete information on a chemical (database uncertainty).

For the *interspecies uncertainty factor*, the default value of 10^1 (10X) is composed of a toxicodynamic (TD) portion of $10^{0.5}$ ($\sim 3.3\times$) and a toxicokinetic (TK) portion of $10^{0.5}$ ($\sim 3.3\times$). This uncertainty factor, in any given assessment, can be adjusted if the available data about interspecies differences in TD and/or TK suggest a similarity or difference between animal models and humans (see Fig. 1). If data on chemical mechanism of action support a similarity or dissimilarity in humans as compared to the animal test model, then the relevant portion of the interspecies uncertainty factor can be adjusted. For example, in the IRIS Tox Review for ethylene glycol monobutyl ether (EGBE), the TD portion of the interspecies uncertainty factor was reduced to 10^0 ($1\times$) based on data indicating that humans are less sensitive than the test animal for the critical effect (U.S. EPA, 2010).

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

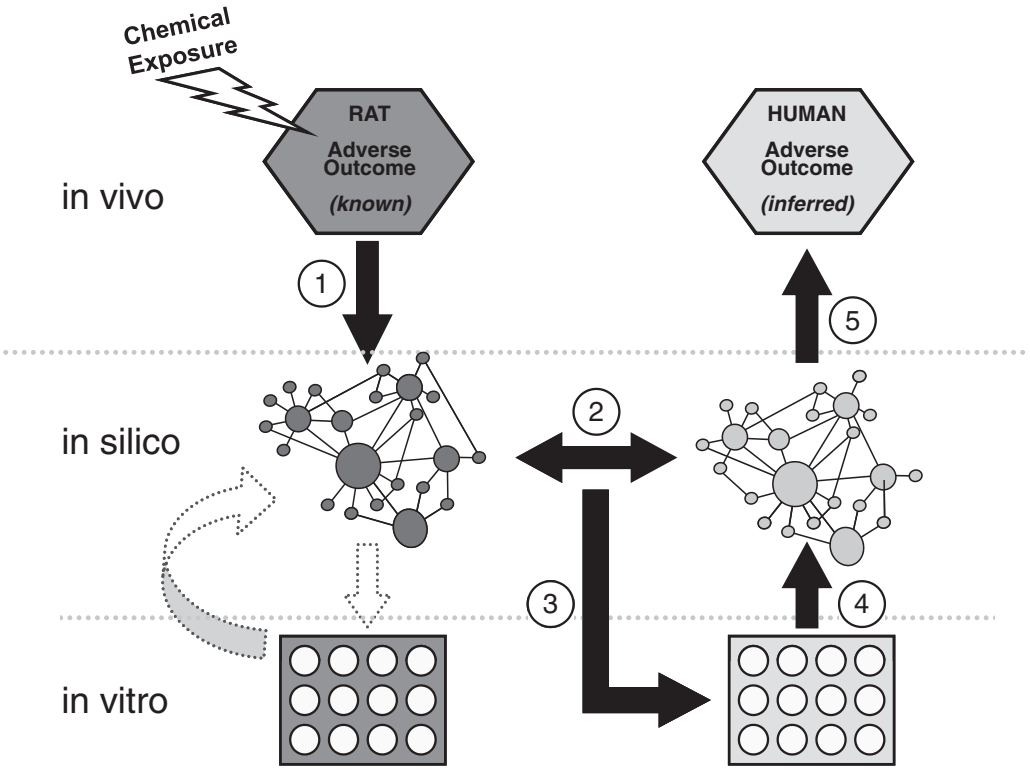

Fig. 3. Interspecies comparisons and human relevance — a modified parallelogram approach for integrating *in vivo*, in vitro, and computational approaches in interspecies extrapolation of toxicity perturbation. The parallelogram approach proposed by Nesnow (2004) and referred to by the National Research Council (National Research Council, 2006) is modified here by the incorporation of computational comparative genomics approaches. Using rat and human as examples: 1) a rat network perturbation model is developed based on *in vivo* data; 2) the rat and human networks are computationally compared; 3) differences and similarities found by the interspecies network comparison are tested via human *in vitro* assays (e.g., primary human cell lines); 4) quantified *in vitro* perturbations are mapped back to the compared networks; and, 5) human *in vivo* outcomes are inferred. In addition, rat *in vivo* assays, driven by network-based hypotheses or otherwise (as represented by the white arrows), can inform the rat network model and the compared network model.

highlights the utility of both gene-level and pathway-level effects as a method to qualitatively identify species-dependent and species-independent gene and pathway effects. With information on multiple species, mechanism of action predictions can be improved for the species of interest.

We conclude that comparative genomics data and methodologies will meet a critical need in environmental chemical risk assessment for understanding interspecies differences and for characterizing the uncertainties resulting from extrapolating inferences across species. To make advancements in chemical risk assessment through better understanding interspecies differences in mechanism of action, we propose that research and development efforts focus on current computational methodologies, as compiled and discussed here (see Table 2 in addition to literature cited in the text), and on sources of publicly available genomic information (see Table 1), that can be utilized and adapted for the purpose of comparing biological systems and processes across species and for defining toxicity-induced responses in species of interest. Specific areas ripe for further research include the following:

- Chemical case studies applying single gene or protein level comparisons to existing risk assessment processes.
- Chemical case studies based on the research paradigm outlined in Fig. 3 and described herein.
- The development of approaches for quantifying the functional changes and physiological implications associated with detected molecular differences across species in pathway comparison approaches.
- Network comparison methods that are shown to be robust to data gaps and that include performance estimates for more quantitative applications.
- Improved characterization of the differences and similarities in pathway and network models of unperturbed systems among species most relevant to human and ecological risk assessment.

In summary, modeling chemical perturbations in test animals onto pathways and networks that can then be compared to human pathways and networks will bring human health risk assessment one step closer to the NRC's 21st century vision of tethering pathway perturbations directly to human disease instead of basing the mechanism of action on animal models. By querying interspecies comparative network and/or pathway models of chemically-induced perturbations and quantifying the differences, it should then be possible to move closer to developing quantitative measures of interspecies differences for use in risk assessment.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

Acknowledgments

We are especially thankful to Dr. Holly Mortensen, of EPA's National Center for Computational Toxicology, for helpful suggestions and critical discussion. We also appreciate the insights provided by Dr. Babasaheb Sonawane of EPA's National Center for Environmental Assessment.

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

References

Albert, R., Jeong, H., Barabasi, A.-L., 2000. Error and attack tolerance of complex networks. Nature 406, 378–382.

Alexeyenko, A., Sonnhammer, E.L.L., 2009. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res. 19, 1107–1116.

Ali, W., Deane, C.M., 2009. Functionally guided alignment of protein interaction networks for module detection. Bioinformatics 25, 3166–3173.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J.P., Chothia, C., Murzin, A.G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226–D229.

Andreeva, A., Prlic, A., Hubbard, T.J.P., Murzin, A.G., 2007. SISYPHUS—structural alignments for proteins with non-trivial relationships. Nucleic Acids Res. 35, D253–D259.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., Murzin, A.G., 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36, D419–D425.

Banks, E., Nabieva, E., Peterson, R., Singh, M., 2008. NetGrep: fast network schema searches in interactomes. Genome Biol. 9, R138.

Barker, D., Pagel, M., 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. 1, e3.

Barker, D., Meade, A., Pagel, M., 2007. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics 23, 14–20.

Barthet, M., Hilu, K., 2008. Evaluating evolutionary constraint on the rapidly evolving gene matK using protein composition. J. Mol. Evol. 66, 85–97.

Benson, W.H., Di Giulio, R.T. (Eds.), 2007. Genomic approaches for cross-species extrapolation in toxicology. Setac Press, Pensacola, FL.

Berg, J., Lässig, M., 2006. Cross-species analysis of biological networks by Bayesian alignment. Proc. Natl. Acad. Sci. 103, 10967–10972.

Berg, J., Lassig, M., Wagner, A., 2004. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol. Biol. 4, 51.

Bergmann, S., Ihmels, J., Barkai, N., 2003. Similarities and differences in genome-wide expression data of six organisms. PLoS Biol. 2, e9.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Beyer, R.P., Fry, R.C., Lasarev, M.R., McConnachie, L.A., Meira, L.B., Palmer, V.S., Powell, C.L., Ross, P.K., Bammler, T.K., Bradford, B.U., Cranson, A.B., Cunningham, M.L., Fannin, R.D., Higgins, G.M., Hurban, P., Kayton, R.J., Kerr, K.F., Kosyk, O., Lobenhofer, E.K., Sieber, S.O., Vliet, P.A., Weis, B.K., Wolfinger, R., Woods, C.G., Freedman, J.H., Linney, E., Kaufmann, W.K., Kavanagh, T.J., Paules, R.S., Rusyn, I., Samson, L.D., Spencer, P.S., Suk, W., Tennant, R.J., Zarbl, H., Members of the Toxicogenomics Research Consortium, 2007. Multicenter study of acetaminophen hepatotoxicity reveals the importance of biological endpoints in genomic analyses. Toxicol. Sci. 99 (1), 326–337.

Binkley, J., Karra, K., Kirby, A., Hosobuchi, M., Stone, E.A., Sidow, A., 2010. ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. Genome Res. 20, 142–154.

Borenstein, E., Kupiec, M., Feldman, M.W., Ruppin, E., 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc. Natl. Acad. Sci. 105, 14482–14487.

Brenner, S.E., Koehl, P., Levitt, M., 2000. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res. 28, 254–256.

Chandonia, J.-M., Walker, N.S., Conte, L.L., Koehl, P., Levitt, M., Brenner, S.E., 2002. ASTRAL compendium enhancements. Nucleic Acids Res. 30, 260–263.

Chandonia, J.-M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S.E., 2004. The ASTRAL compendium in 2004. Nucleic Acids Res. 32, D189–D192.

Chen, M., Hofestadt, R., 2004. PathAligner: metabolic pathway retrieval alignment. Appl. Bioinforma. 3, 241–252.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., Leonardson, A., Castellini, L.W., Wang, S., Champy, M.-F., Zhang, B., Emilsson, V., Doss, S., Ghazalpour, A., Horvath, S., Drake, T.A., Lusis, A.J., Schadt, E.E., 2008. Variations in DNA elucidate molecular networks that cause disease. Nature 452, 429–435.

Clemente, J.C., Satou, K., Valiente, G., 2005. Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. Genome Inform. 16, 45–55.

Clemente, J.C., Satou, K., Valiente, G., 2007. Phylogenetic reconstruction from non-genomic data. Bioinformatics 23, e110–e115.

Cokus, S., Mizutani, S., Pellegrini, M., 2007. An improved method for identifying functionally linked proteins using phylogenetic profiles. BMC Bioinforma. 8, S7.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., Sidow, A., 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15, 901–913.

Cordero, O.X., Snel, B., Hogeweg, P., 2008. Coevolution of gene families in prokaryotes. Genome Res. 18, 462–468.

Costa, L.d.F., Rodrigues, F.A., Cristino, A.S., 2008. Complex networks: the key to systems biology. Genet. Mol. Biol. 31, 591–601.

Date, S.V., Marcotte, E.M., 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat. Biotechnol. 21, 1055–1062.

Deville, Y., Gilbert, D., van Helden, J., Wodak, S.J., 2003. An overview of data models for the analysis of biochemical pathways. Brief. Bioinform. 4, 246–259.

Dutkowski, J., Tiuryn, J., 2007. Identification of functional modules from conserved ancestral protein protein interactions. Bioinformatics 23, i149–i158.

Dutkowski, J., Tiuryn, J., 2009. Phylogeny-guided interaction mapping in seven eukaryotes. BMC Bioinforma. 10, 393.

Edwards, S.W., Preston, R.J., 2008. Systems biology and mode of action based risk assessment. Toxicol. Sci. 106, 312–318.

Erten, S., Li, X., Bebek, G., Li, J., Koyuturk, M., 2009. Phylogenetic analysis of modularity in protein interaction networks. BMC Bioinforma. 10, 333.

Falicov, A., Cohen, F.E., 1996. A surface of minimum area metric for the structural comparison of proteins. J. Mol. Biol. 258, 871–892.

Flannick, J., Novak, A., Srinivasan, B.S., McAdams, H.H., Batzoglou, S., 2006. Græmlin: general and robust alignment of multiple large interaction networks. Genome Res. 16, 1169–1181.

Flannick, J., Novak, A., Do, C.B., Srinivasan, B.S., Batzoglou, S., 2009. Automatic parameter learning for multiple local network alignment. J. Comput. Biol. 16, 1001–1022.

Forst, C.V., Schulten, K., 1999. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. J. Comput. Biol. 6, 343–360.

Forst, C.V., Schulten, K., 2001. Phylogenetic analysis of metabolic pathways. J. Mol. Evol. 52, 471–489.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., Hardison, R.C., 2003. Cross-species sequence comparisons: a review of methods and available resources. Genome Res. 13, 1–12.

Gonzalez, O., Zimmer, R., 2008. Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. Bioinformatics 24, 1257–1263.

Goode, D.L., Cooper, G.M., Schmutz, J., Dickson, M., Gonzales, E., Tsai, M., Karra, K., Davydov, E., Batzoglou, S., Myers, R.M., Sidow, A., 2010. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Res. 20, 301–310.

Gough, J., 2005. Convergent evolution of domain architectures (is rare). Bioinformatics 21, 1464–1471.

Gunnarsson, L., Jauhiainen, A., Kristiansson, E., Nerman, O., Larsson, D.G.J., 2008. Evolutionary conservation of human drug targets in organisms used for environmental risk assessments. Environ. Sci. Technol. 42, 5807–5813.

Guo, X., Hartemink, A.J., 2009. Domain-oriented edge-based alignment of protein interaction networks. Bioinformatics 25, i240–i1246.

Han, J.-D.J., 2008. Understanding biological functions through molecular networks. Cell Res. 18, 224–237.

Hartig, P.C., Cardon, M.C., Lambright, C.R., Bobseine, K.L., Gray, L.E., Wilson, V.S., 2007. Substitution of synthetic chimpanzee androgen receptor for human androgen receptor in competitive binding and transcriptional activation assays for EDC screening. Toxicol. Lett. 174, 89–97.

Hennig, W., 1965. Phylogenetic systematics. Annu. Rev. Entomol. 10, 97–116.

Heymans, M., Singh, A.K., 2003. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics 19, i138–i146.

Hotchkiss, A.K., Ostby, J.S., Vandenbergh, J.G., 2002. Androgens and environmental antiandrogens affect reproductive development and play behavior in the Sprague–Dawley rat. Environ. Heal. Perspect. 110, 435–439.

Huang, C.C., Novak, W.R., C, B. P., Jewett, A.I., Ferrin, T.E., Klein, T.E., 2000. Integrated tools for structural and sequence alignment and analysis. Pac. Symp. Biocomput. 2000, 230–241.

Jewett, A.I., Huang, C.C., Ferrin, T.E., 2003. MINRMS: an efficient algorithm for determining protein structure similarity using root-mean-squared-distance. Bioinformatics 19, 625–634.

Joshi, A., Van Parys, T., Peer, Y., Michoel, T., 2010. Characterizing regulatory path motifs in integrated networks using perturbational data. Genome Biol. 11, R32.

Jothi, R., Przytycka, T., Aravind, L., 2007. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinforma. 8, 173.

Kalaev, M., Smoot, M., Ideker, T., Sharan, R., 2008. NetworkBLAST: comparative analysis of protein networks. Bioinformatics 24, 594–596.

Kastenmuller, G., Schenk, M., Gasteiger, J., Mewes, H.-W., 2009. Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. Genome Biol. 10, R28.

Kawabata, T., 2003. MATRAS: a program for protein 3D structure comparison. Nucleic Acids Res. 31, 3367–3369.

Kawabata, T., Nishikawa, K., 2000. Protein structure comparison using the Markov transition model of evolution. Proteins: Struct. Funct. Genet. 41, 108–122.

Kelley, R., Ideker, T., 2005. Systematic interpretation of genetic interactions using protein networks. Nat. Biotechnol. 23, 561–566.

Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., Ideker, T., 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. U. S. A. 100, 11394–11399.

Kholodenko, B.N., 2006. Cell-signalling dynamics in time and space. Nat. Rev. Mol. Cell Biol. 7, 165–176.

Klau, G., 2009. A new graph-based method for pairwise global network alignment. BMC Bioinforma. 10, S59.

Koyuturk, M., Grama, A., Szpankowski, W., 2004. An efficient algorithm for detecting frequent subgraphs in biological networks. Bioinformatics 20, i200–i207.

Koyuturk, M., Kim, Y., Topkara, U., Subramaniam, S., Szpankowski, W., Grama, A., 2006. Pairwise alignment of protein interaction networks. J. Comput. Biol. 13, 182–199.

Kumar, P., Henikoff, S., Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–1082.

LeCluyse, E.L., 2001. Pregnane X receptor: molecular basis for species differences in CYP3A induction by xenobiotics. Chem. Biol. Interact. 134, 283–289.

Li, Y., de Ridder, D., de Groot, M., Reinders, M., 2008. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. BMC Syst. Biol. 2, 111.

Liao, L., Noble, W.S., 2002. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. Proceedings of The Sixth International Conference on Research in Computational Molecular Biology, pp. 225–232.

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

Liao, C.-S., Lu, K., Baym, M., Singh, R., Berger, B., 2009. IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics 25, i253–i258.
Lin, J., Qian, J., Greenbaum, D., Bertone, P., Das, R., Echols, N., Senes, A., Stenger, B., Gerstein, M., 2002. GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. Nucleic Acids Res. 30, 4574–4582.
Linghu, B., Snitkin, E., Holloway, D., Gustafson, A., Xia, Y., DeLisi, C., 2008. High-precision high-coverage functional inference from integrated data sources. BMC Bioinforma. 9, 119.
Lo Conte, L., Ailey, B., Hubbard, T.J.P., Brenner, S.E., Murzin, A.G., Chothia, C., 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res. 28, 257–259.
Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J.I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J.B., Bickel, P., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Stone, E.A., Rosenbloom, K.R., Kent, W.J., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V.B., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Hinrichs, A., Trumbower, H., Clawson, H., Zweig, A., Kuhn, R.M., Barber, G., Harte, R., Karolchik, D., Field, M.A., Moore, R.A., Matthewson, C.A., Schein, J.E., Marra, M.A., Antonarakis, S.E., Batzoglou, S., Goldman, N., Hardison, R., Haussler, D., Miller, W., Pachter, L., Green, E.D., Sidow, A., 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 17, 760–774.
Mattes, W.B., 2006. Cross-species comparative toxicogenomics as an aid to safety assessment. Expert Opin. Drug Metab. Toxicol. 2, 859–874.
Mazurie, A., Bonchev, D., Schwikowski, B., Buck, G.A., 2008. Phylogenetic distances are encoded in networks of interacting pathways. Bioinformatics 24, 2579–2585.
Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H., Thomas, P.D., 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 33, D284–D288.
Monroe, D., 2009. Genomic Clues to DNA Treasure Sometimes Lead Nowhere. Science 325, 142–143.
Moore, L.B., Maglich, J.M., McKee, D.D., Wisely, B., Willson, T.M., Kliewer, S.A., Lambert, M.H., Moore, J.T., 2002. Pregnane X receptor (PXR), constitutive androstane receptor (CAR), and benzoate X receptor (BXR) define three pharmacologically distinct classes of nuclear receptors. Mol. Endocrinol. 16, 977–986.
Mortensen, H.M., Euling, S.Y., this issue. Integrating mechanistic and polymorphism data to characterize human genetic susceptibility for environmental chemical risk assessment in the 21st century. *Toxicology and Applied Pharmacology* **Genomics and Risk Assessment Special Issue**.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Narayanan, M., Karp, R.M., 2007. Comparing protein interaction networks via a graph match-and-split algorithm. J. Comput. Biol. 14, 892–907.
National Research Council, 2006. Application of Toxicogenomics to Cross-Species Extrapolation. The National Academies Press, Washington, DC.
National Research Council, 2007. Toxicity Testing in the 21st Century: a Vision and a Strategy. The National Academies, Washington, DC.
National Research Council, 2009. Science and Decisions: Advancing Risk Assessment. The National Academies Press, Washington, DC.
Nesnow, S., 2004. A transcriptional analysis approach to understanding the basis of species differences in conazole toxicology. Presentation at the Seventh Meeting on Emerging Issues and Data on Environmental Contaminants - Applications of Toxicogenomics to Cross Species Extrapolation: A Workshop, August 12, 2004, Washington, DC.
Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814.
Nielsen, R. (Ed.), 2005. Statistics for Biology and Health. Springer Press.
Ovacik, M.A., Androulakis, I.P., this issue. *Toxicology and Applied Pharmacology*.
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U. S. A. 96, 4285–4288.
Pettit, S., des Etages, S.A., Mylecraine, L., Snyder, R., Fostel, J., Dunn, R.T., Haymes, K., Duval, M., Stevens, J., Afshari, C., Vickers, A., 2010. Current and future applications of toxicogenomics: results summary of a survey from the HESI genomics state of science subcommittee. Environ. Health Perspect. 118, 992–997.
Pinney, J.W., Amoutzias, G.D., Rattray, M., Robertson, D.L., 2007. Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. Proc. Natl. Acad. Sci. 104, 20449–20453.
Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E., Ziv-Ukelson, M., 2005. Alignment of metabolic pathways. Bioinformatics 21, 3401–3408.
Qian, X., Yoon, B.-J., 2009. Effective identification of conserved pathways in biological networks using hidden Markov models. PLoS One 4, e8070.
Raghava, G., Barton, G., 2006. Quantification of the variation in percentage identity for protein sequence alignments. BMC Bioinforma. 7, 415.
Rakhshandehroo, M., Hooiveld, G., Müller, M., Kersten, S., 2009. Comparative analysis of gene regulation by the transcription factor PPARalpha between mouse and human. PLoS One 4 (8), e6796.
Ranea, J.A.G., Yeats, C., Grant, A., Orengo, C.A., 2007. Predicting protein function with hierarchical phylogenetic profiles: the gene3d phylo-tuner method applied to eukaryotic genomes. PLoS Comput. Biol. 3, e237.
Rivera, C., Vakil, R., Bader, J., 2010. NeMo: network module identification in cytoscape. BMC Bioinforma. 11, S61.
Ruano-Rubio, V., Poch, O., Thompson, J., 2009. Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. BMC Bioinforma. 10, 383.
Schadt, E.E., 2009. Molecular networks as sensors and drivers of common human diseases. Nature 461, 218–223.
Searls, D.B., 2003. Pharmacophylogenomics: genes, evolution and drug targets. Nat. Rev. Drug Discov. 2, 613–623.
Shah, I.A., Wambaugh, J.F., 2010. Virtual tissues in toxicology. J. Toxicol. Environ. Health B Crit. Rev. 13, 314–328.
Sharan, R., Ideker, T., 2006. Modeling cellular machinery through biological network comparison. Nat. Biotechnol. 24, 427–433.
Sharan, R., Ideker, T., Kelley, B.P., Shamir, R., Karp, R.M., 2005a. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. J. Comput. Biol. 12, 835–846.
Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., Ideker, T., 2005b. Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. U. S. A. 102, 1974–1979.
Sharan, R., Ulitsky, I., Shamir, R., 2007. Network-based prediction of protein function. Mol. Syst. Biol. 3, 1–13.
Shindyalov, I., Bourne, P., 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739–747.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.
Simon, A.L., Stone, E.A., Sidow, A., 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. Proc. Natl. Acad. Sci. 99, 2912–2917.
Singh, R., Xu, J., Berger, B., 2008. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. Natl. Acad. Sci. 105, 12763–12768.
Stone, E.A., Sidow, A., 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 15, 978–986.
Stuart, J.M., Segal, E., Koller, D., Kim, S.K., 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255.
Sun, Y.V., Boverhof, D.R., Burgoon, L.D., Fielden, M.R., Zacharewski, T.R., 2004. Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. Nucleic Acids Res. 32 (15), 4512–4523.
Tan, C.S.H., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M.O., Jorgensen, C., Bader, G.D., Aebersold, R., Pawson, T., Linding, R., 2009. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. Sci. Signal. 2, ra39-.
Taylor, W.R., Orengo, C.A., 1989. Protein structure alignment. J. Mol. Biol. 208, 1–22.
Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A., 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 13, 2129–2141.
Tian, W., Samatova, N.F., 2009. Pairwise alignment of interaction networks by fast identification of maximal conserved patterns. Proceedings of the Pacific Symposium on Biocomputing, pp. 99–110.
Tohsato, Y., Matsuda, H., Hashimoto, A., 2000. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, pp. 376–383.
Tun, K., Dhar, P., Palumbo, M., Giuliani, A., 2006. Metabolic pathways variability and sequence/networks comparisons. BMC Bioinforma. 7, 24.
U.S. EPA, 1998. Guidelines for Neurotoxicity Risk Assessment. U.S. Environmental Protection Agency, Washington, DC.
U.S. EPA, 2006. A Framework for Assessing Health Risk of Environmental Exposures to Children (Final). U.S. Environmental Protection Agency, Washington, DC.
U.S. EPA, 2009. The U.S. Environmental Protection Agency's Strategic Plan for Evaluating the Toxicity of Chemicals. U.S. Environmental Protection Agency.
U.S. EPA, 2010. Toxicological review of ethylene glycol monobutyl ether (EGBE) (Cas No. 111-76-2) (Final). In: Integrated Risk Information System (Ed.), U.S. Environmental Protection Agency, Washington, DC.
Uehara, T., Ono, A., Maruyama, T., Kato, I., Yamada, H., Ohno, Y., Urushidani, T., 2010. The Japanese toxicogenomics project: application of toxicogenomics. Mol. Nutr. Food Res. 54 (2), 218–227.
Wang, Z., Zhang, J., 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? PLoS Genet. 5, e1000329.
Wang, K., Narayanan, M., Zhong, H., Tompa, M., Schadt, E.E., Zhu, J., 2009. Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. PLoS Comput. Biol. 5, e1000616.
Wiles, A., Doderer, M., Ruan, J., Gu, T.-T., Ravi, D., Blackman, B., Bishop, A., 2010. Building and analyzing protein interactome networks by cross-species comparisons. BMC Syst. Biol. 4, 36.
Wilson, V.S., Cardon, M.C., Thornton, J., Korte, J.J., Ankley, G.T., Welch, J., Gray, L.E., Hartig, P.C., 2004. Cloning and in vitro expression and characterization of the androgen receptor and isolation of estrogen receptor α from the fathead minnow (Pimephales promelas). Environ. Sci. Technol. 38, 6314–6321.
Wilson, V.S., Cardon, M.C., Gray, L.E., Hartig, P.C., 2007. Competitive binding comparison of endocrine-disrupting compounds to recombinant androgen receptor from fathead minnow, rainbow trout, and human. Environ. Toxicol. Chem. 26, 1793–1802.

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011

Wu, J., Kasif, S., DeLisi, C., 2003. Identification of functional links between genes using phylogenetic profiles. Bioinformatics 19, 1524–1530.
Wuchty, S., Ravasz, E., Barabási, A.-L., 2006. The architecture of biological networks. In: Deisboeck, T.S., Kresh, J.Y. (Eds.), Complex Systems Science in Biomedicine. Springer, New York, pp. 165–181.
Yamada, T., Kanehisa, M., Goto, S., 2006. Extraction of phylogenetic network modules from the metabolic network. BMC Bioinforma. 7, 130.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford.
Zaslavskiy, M., Bach, F., Vert, J.-P., 2009. Global alignment of protein–protein interaction networks by graph matching methods. Bioinformatics 25, i259–i1267.
Zhou, Y., Wang, R., Li, L., Xia, X., Sun, Z., 2006. Inferring functional linkages between proteins from evolutionary scenarios. Journal of Molecular Biology 359, 1150–1159.
Zhu, X., Gerstein, M., Snyder, M., 2007. Getting connected: analysis and principles of biological networks. Genes & Development 21, 1010–1024.
Zmasek, C., Eddy, S., 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3, 14.

Please cite this article as: Burgess-Herbert, S.L., Euling, S.Y., Use of comparative genomics approaches to characterize interspecies differences in response to environmental ..., Toxicol. Appl. Pharmacol. (2011), doi:10.1016/j.taap.2011.11.011