The Unreasonable Effectiveness of Inverse Reinforcement Learning in Advancing Cancer Research

John Kalantari,[1,2] **Heidi Nelson,**[1,3] **Nicholas Chia**[1,2,4]
[1]Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA
[2]Division of Surgical Research, Department of Surgery, Mayo Clinic, Rochester, MN, USA
[3]Division of Colon and Rectal Surgery, Department of Surgery, Mayo Clinic, Rochester, MN, USA
[4]Department of Physiology and Biomedical Engineering, Mayo Clinic, Rochester, MN, USA
{kalantari.john, chia.nicholas}@mayo.edu

Abstract

The “No Free Lunch” theorem states that for any algorithm, elevated performance over one class of problems is offset by its performance over another. Stated differently, no algorithm works for everything. Instead, designing effective algorithms often means exploiting prior knowledge of data relationships specific to a given problem. This “unreasonable efficacy” is especially desirable for complex and seemingly intractable problems in the natural sciences. One such area that is rife with the need for better algorithms is cancer biology—a field where relatively few insights are being generated from relatively large amounts of data. In part, this is due to the inability of mere statistics to reflect cancer as a genetic evolutionary process—one that involves cells actively mutating in order to navigate host barriers, outcompete neighboring cells, and expand spatially.

Our work is built upon the central proposition that the Markov Decision Process (MDP) can better represent the process by which cancer arises and progresses. More specifically, by encoding a cancer cell’s complex behavior as a MDP, we seek to model the series of genetic changes, or evolutionary trajectory, that leads to cancer as an optimal decision process. We posit that using an Inverse Reinforcement Learning (IRL) approach will enable us to reverse engineer an optimal policy and reward function based on a set of “expert demonstrations” extracted from the DNA of patient tumors. The inferred reward function and optimal policy can subsequently be used to extrapolate the evolutionary trajectory of any tumor. Here, we introduce a Bayesian nonparametric IRL model (PUR-IRL) where the number of reward functions is *a priori* unbounded in order to account for uncertainty in cancer data, i.e., the existence of latent trajectories and non-uniform sampling. We show that PUR-IRL is “unreasonably effective” in gaining interpretable and intuitive insights about cancer progression from high-dimensional genome data.

1 Introduction

In the follow up survey paper (Halevy, Norvig, and Pereira 2009) to ”The Unreasonable Effectiveness of Mathematics in the Natural Sciences” (Wigner 1960) it was argued that solving the most complex AI problems require embracing the unreasonable effectiveness of data. That is, solving the world’s most challenging problems requires we take advantage of the complex structure and relationships inherent in real-world data. This important insight has already been intuited by AI communities working in computer vision and natural language processing, resulting in improved AI solutions (Li et al. 2019; Wang and Wan 2018). Unfortunately, the majority of cancer research exemplifies what happens when one does the opposite and uses statistical approaches that assume no meaningful structure exists in cancer data. Historically, this occurred because there was no algorithmic means to account for the complexities inherent to cancer and because our understanding of cancer as a complex process was lacking. Here, we demonstrate the impact of considering the underlying biological processes of cancer evolution into the algorithmic design of tools for studying cancer progression. More specifically, we argue that Inverse Reinforcement Learning (IRL) is an unreasonably effective algorithm for gaining interpretable and intuitive insight about cancer progression because of its ability to take advantage of prior knowledge about the structure and source of its input data.

In support of this, we implement the **Pop-Up Restaurant for Inverse Reinforcement Learning** (**PUR-IRL**) approach—a Bayesian nonparametric IRL model that takes advantage of relationships between events in seemingly disparate data sources by allowing for the inference of multiple reward functions from non-uniformly sampled data. In testing PUR-IRL on real-world data from colorectal cancer (CRC) patients, we verify its ability to infer a series of mutational events, or an evolutionary trajectory, that broadly matches those arrived at by CRC experts through the curation of a variety of multi-omics and experimental data sources. Furthermore, we show that PUR-IRL can accomplish this with data taken from a mere tens of patients and that this outperforms frequency-based statistical approaches that are commonly used in cancer research. Our experimental results show that PUR-IRL can correctly identify the number of distinct experts, the reward function and optimal policy of each expert, and remain robust in classification under various data sampling conditions. Tested on GridWorld, PUR-IRL achieved an F1-score of 0.9328 and 0.90331 under uniform and non-uniform sampling conditions, respectively.

2 Motivation

Cancer is a complex decision-making process. It is a population dynamic process whereby choices are made through the sequential accumulation of advantageous genetic changes—i.e., evolution. Long mimicked computationally (Koza 1992), evolution represents an optimization process whereby success is rewarded and failure is erased. From the myopic viewpoint of a single cell in the human body, success means survival and replication. These are the two evolutionary forces that bias a normal healthy cell towards a state of uncontrolled growth that is typically associated with cancer. Fortunately, a cell's evolution to cancer does not happen overnight. There exist a complex genetic and regulatory maze of barriers that prevent cells from becoming cancerous. This is why humans, for the most part, do not get cancer in the earlier part of their lives. Only after the evolutionary game has played out a large multitude of times (i.e. through many cellular replications), do the majority of humans eventually succumb to cancer. Indeed, cancer progression is a more nuanced process, enabled by the acquisition of genetic mutations that allow for a cell to gradually overcome human-host defense mechanisms. In essence, expert navigation is required by cancer in order to evade the immune system, overgrow neighboring cells, and expand beyond the normal spatial compartments.

In order to truly conquer cancer for future patients, we need to understand how cancer overcomes our host-defenses from a unified mechanistic perspective. Myriad observational studies have taught us that we cannot merely wait for the same thing to happen twice because no two cancers are exactly alike. They occur in patients with different genetic backgrounds and accumulate different genetic alterations. Yet, despite these differences, they are unified by similar mechanisms or types of genetic changes. In other words, there are multiple etiological paths tied together by specific events that share commonality in their causal mechanism. Understanding these common mechanisms will enable the development of better therapies and preventative measures. It will also enable improved prediction of recurrence and metastatic advancement of cancer, directly impacting the 606,880 annual cancer deaths in the United States alone (Siegel, Miller, and Jemal 2019).

Our work is motivated by the multiple parallels between inverse reinforcement learning (IRL) algorithms and cancer evolution. First, reinforcement learning allows us to encode the complex behavior of an independent agent as a Markov Decision Process (MDP). If an agent is known to behave 'optimally', as can be assumed for cancer, there exists an optimal policy and an underlying reward function. This reward function structures the space of possible policies that make up the solutions to the MDP and are capable of reflecting the multifaceted nature of cancer. Secondly, by extending Markov chains with the addition of 'actions' and 'rewards' to model choice and indicate preference, the MDP becomes suitable for modeling sequential decision-making processes. The use of a finite state-action space within a stochastic environment makes MDP an interpretable modeling paradigm for encoding the genetic alterations that take place within a large, combinatorial event space. Third, the probabilistic nature of the MDP allows us to cope with imperfect data. Randomness in cancer data is both intrinsic to the stochastic nature of the evolutionary process as well as extrinsically imposed by the ethical limitations and practice-based trade-offs of medicine. For example, one can often only sample piecemeal from tumor sections that are not needed for clinical purposes, therefore biasing the sampling procedure. In addition, there are ethical bounds to observing the progression of a tumor or precursor lesion when excision and/or treatment are in the best interests of the patient. A probabilistic approach to reasoning about uncertainty can be taken during the IRL process in order to account for demonstrated behavior that is prone to noise (Ziebart et al. 2008), data that may have been collected from multiple agents (Choi and Kim 2012), or any other uncertainty that may still exist due to the structure and source of the observed optimal decision process. Fourth, the IRL problem reduces to recovering a reward function that induces the demonstrated behavior with a search algorithm to enforce consistency among the state-action pairs observed in the expert demonstrations. Once the latent reward function describing the explicit values of various state and action pairs, and optimal policy defining the general (non-surjective and non-injective) mapping from states to actions are inferred, implicit causal relationships encoded within the data can be extrapolated for subsequent predictive and mechanistic modeling tasks. By assuming the near-optimal behavior of cancer, the reward weights of this reward function can be inferred from examples of the agent's behavior using IRL (Abbeel and Ng 2004). This has the added advantage of closely matching how cancer data needs to be used in real-world applications, since one cannot ethically watch cancer progress unchecked in a patient without immediate intervention. For these reasons, IRL closely parallels both the nature of the underlying evolutionary process and realities of the cancer data, making IRL ideal for modeling cancer as a complex decision-making process. The successful prevention and treatment of cancer (Burrell et al. 2013; Schwartz and Schäffer 2017) requires that we distill knowledge from ambiguous and problematic data, without direct experimental validation or a gold-standard. The solutions provided need to be mappable to biological mechanisms, i.e., biologically interpretable. Finally, the algorithm must be both computable and accurate. Because there is no direct validation possible, these must be shown for simulated data. In this paper, we use real and simulated data to demonstrate that

- Reward functions resolve some of the inherent problems with tumor data, mainly, tumor heterogeneity.
- PUR-IRL usefully captures uncertainties within the cancer data and arrives at biologically relevant conclusions.
- PUR-IRL allows for incremental integration of new information through iterative updates, thereby turning one large intractable problem into a series of tractable ones.
- PUR-IRL accurately infers optimal policies and latent reward functions given a set of expert demonstrations.

3 Methods

3.1 Data Pre-processing

Raw Data Generation. Whole genome sequencing (WGS) was performed on samples from a previously described study (Hale et al. 2018a; 2018b) In brief, samples consist of normal and tumor tissue pairs from 27 patients. Sequencing was performed using the BGISEQ-500 (2x100bp kit, $\tilde{3}$0x) and data reads were mapped to human reference genome GRCh38 with decoy sequences (Li and Durbin 2009). Somatic mutations and indels were determined by comparing tumor samples with normal samples using MuTect2 and subsequently filtered using FilterMutectCalls from the Genome Analysis Toolkit (GATK) (DePristo et al. 2011). Aneuploidy and somatic copy number alterations were determined using Titan (Ha et al. 2014) and used to infer sample purity. Variant annotation was performed using SNPeff (Cingolani et al. 2012). Data available upon request.

Extracting expert demonstrations of cancer progression from patient tumors. Tumors are comprised of multiple genetically diverse subclonal populations of cells, each harboring distinct mutations. While different subclones can appear distinct, prior knowledge tells us that they are related to one another through the process of evolution, i.e., the sequential acquisition of random mutations (Valastyan and Weinberg 2011). Using this prior knowledge, the evolutionary relationship between these subclonal populations can be described in a series of linear and branching evolutionary expansions and modeled as a phylogenetic tree. One can assume that a cancer cell, which may exist as one of N subclones, has undergone a sequence of alterations that serve to maximize a set of rewards (i.e., growth and survival) within a competitive environment where the neighboring cancer subpopulations are competing for resources (Valastyan and Weinberg 2011; Schwartz and Schäffer 2017). The distinct sequence of subclones visited while traversing down from the root node down to a leaf node of a tumor's phylogenetic tree (Fig. 1D) can be considered a *path* or expert-demonstration of a cancer subclone's optimal behavior and serve as the input to the PUR-IRL algorithm.

In other words, the cells present from the tumor are the result of competition and selection. Cancer cells that underwent alterations that do not provide a competitive advantage will be outcompeted. These sub-optimal cells will likely not survive nor be observed in the tumor biopsy. Conversely, the most abundant cancer cells found in tumor can be considered the winners of the evolutionary game by having collected a combination of mutations beneficial to replication and population expansion—the properties of cancer. If all cancer cells within a tumor can be identified as corresponding to one of only a finite number of subclone profiles (out of millions of possibilities), we know that the sequence of actions displayed by all cells with the same subclone profile were optimal for the survival and growth of that tumor.

The field of tumor phylogenetics encompasses a variety of techniques focused on the problem of subclonal reconstruction (Schwartz and Schäffer 2017; El-Kebir et al. 2015). The primary focus of such algorithms has been the deconvolution of genomic data from an observed tumor into its constituent subclones. In general, we are not given the somatic mutations for each tumor subclone. Instead, we have to infer these based on the variant allele fractions (VAFs) from bulk sequencing, i.e., the sum of mutations from all subclones within that sample. These subclonal mutations are then used to determine the phylogenetic relationships between subclones. However, these techniques have two key limitations. First, they almost never produce a unique solution. That is, for any set of genomic data extracted from a tumor, there will typically exist multiple, equally valid solutions (phylogenetic trees) (Deshwar et al. 2015). Secondly, these techniques do not provide a framework for uniting disparate observations from separate tumors into a general model for understanding the drivers of carcinogenesis.

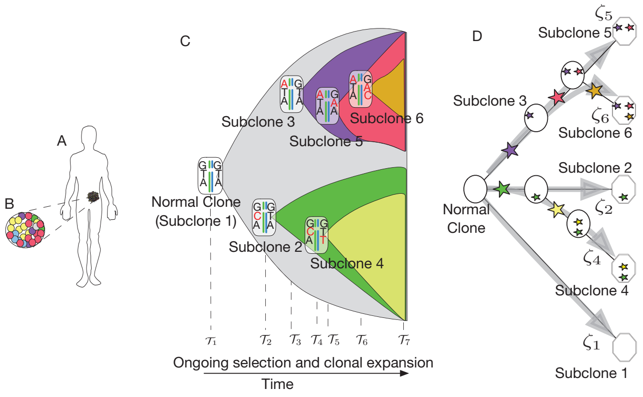

Figure 1: Tracking Tumor Evolution and Reconstructing a Model of Clonal Architecture. Figure 1A-1B: Illustration of a biopsy sample from patient colon tumor. Figure 1C: A model of clonal evolution based on the median values for the VAFs from each inferred subclonal population. Figure 1D: Representative phylogenetic tree of clonal architecture based on biopsy sample.

IRL methods such as PUR-IRL embrace the combinatorial explosion of paths by which each subclonal population of cancer cells may have developed by trying to unite under a single optimal policy specifying the 'general rules' by which cancer progresses and a reward function elucidating how the set of diverse set of state-action pairs observed across subclonal demonstrations are related.

3.2 Pop-Up Restaurant Process for Inverse Reinforcement Learning

Overview of IRL. Inverse Reinforcement Learning (IRL) infers an environment's reward function given observations of an optimally-behaving agent (Abbeel and Ng 2004; Ng and Russell 2000). Such problems can be modeled using a mathematical framework for sequential decision-making known as the *Markov Decision Process* (MDP). This model is defined in terms of a set of *states* S; a set of *actions* A; a stochastic transition distribution $P(s_{t+1}|a_t, s_t)$, describing the probability of outcomes following the execution of an action a_t in state s_t; and a reward function $R(s_t, a_t)$.

Given an $MDP\backslash R$, inverse reinforcement learning identifies a reward function R under which π^* matches the paths $\zeta = \{\zeta_1, \zeta_2, ..., \zeta_M\}$, where ζ_i is a sequence of state-action pairs $\zeta_i = \{(s_{i,0}, a_{i,0}), (s_{i,1}, a_{i,1}), ..., (s_{i,T}, a_{i,T})\}$. In many cases, this observed behavior can be given explicitly as an optimal policy π^* or as a set of sample paths generated by an agent following π^*.

PUR-IRL:Embracing Uncertainty during IRL. Here, we describe a general-purpose and data-agnostic algorithm called the Pop-Up Restaurant Process for Inverse Reinforcement Learning (PUR-IRL) which can infer multiple latent reward functions from a set of expert demonstrations and use these to adapt the MDP architecture in order to integrate novel data types. The name of this algorithm alludes to the periodic updating of the MDP architecture used by the Chinese Restaurant Process (CRP). Within each periodic update, a new 'pop-up' CRP is used for the purpose of sampling and partitioning expert demonstrations among K MDP's, each of which with its own latent reward function r_k. The CRP is a computationally tractable metaphor of the Polya urn scheme (Johnson and Kotz 1977) that uses the following analogy: consider a Chinese restaurant with an unbounded number of tables. An observation, x_i, corresponds to a customer entering the restaurant, and the distinct values z_k^* correspond to the tables at which customers can sit. Assuming an initially empty restaurant, the CRP is expressed:

- With probability proportional to $c_{i-1}^{z_k^*} - d$, the i-th customer sits at the table indexed by z_k^*, in which case $x_i = z_k^*$, where $c_{i-1}^{z_k^*}$ denotes the total number of customers sitting at a table with distinct value z_k^* and d is a scalar discount parameter

- With probability proportional to $\alpha + Kd$, the i-th customer sits at a new table, in which case $x_i \sim H$, where α is a scalar concentration parameter , K is the total number of tables, and H is a random probability measure.

By using the CRP, where a Bayesian nonparametric prior represents all variables and how they relate to the data, we can better resolve multiple probabilistic paths to cancer. In addition, the Bayesian nature of the CRP allows us to work naturally with the uncertainty of the underlying data as well as the highly skewed prevalence of events and paths in cancer patients. By applying the CRP within the IRL paradigm, we can learn K reward functions, as $K \rightarrow \infty$, from a set of data paths inferred by tumor phylogenetics.

Bayesian Nonparametric Priors in PUR-IRL. The probabilistic approach taken by PUR-IRL is similar to a previously described Bayesian nonparametric method known as Dirichlet Process Mixture Inverse Reinforcement Learning (DPM-BIRL) (Choi and Kim 2012). Both methodologies share the notion of applying a prior on each of the reward functions $\hat{r}_{t_k}$ to encode preference and a likelihood to measure the compatibility of the reward function with the data, with PUR-IRL utilizing the Pitman-Yor Process (PYP) and an additional discount parameter $d \in [0, 1)$, where $d = 0$ reduces the model to a Dirichlet Process. Together, α and d control the formation of new reward functions.

A key property of any model based on Dirichlet or Pitman-Yor processes is that the posterior distribution provides a partition of the data into clusters, without requiring that the number of clusters be specified in advance. However, this form of Bayesian clustering imposes an implicit *a priori* "rich get richer" property, leading to partitions consisting of a small number of large clusters. To combat this, the use of discount parameter d is used to reduce the probability of adding a new observation to an existing cluster. The PYP prior is particularly well-suited for multi-reward function IRL applications where the set of expert-demonstrations generated by the various ground-truth reward functions may not follow a uniform distribution. The purpose of extending the IRL to use this stochastic process is to control the power-law property via the discount parameter which can induce a long-tail phenomena of a distribution.

Generative Model. In PUR-IRL, the likelihood is defined as an exponential distribution that utilizes the optimal Q-function computed using reward function $\hat{r}$ and an inverse temperature parameter η that governs the exploration-exploitation tradeoff (small $\eta > 0$ represents large noise, all actions are equally probable; large η represents small noise and more greedy policy):

$$P(\zeta|\hat{r}, \eta) = \prod_{m=1}^{M} \prod_{n=1}^{N} P(a_{c_m,n}|s_{c_m,n}\hat{r}, \eta) \tag{1}$$

$$= \prod_{m=1}^{M} \prod_{n=1}^{N} \frac{e^{(\eta Q^*(s_{c_m,n}, a_{c_m,n}; \hat{r}))}}{\sum_{a'} e^{(\eta Q^*(s_{c_m,n}, a'; \hat{r}))}} \tag{2}$$

The posterior distribution can then be given by Bayes' theorem as $\overset{posterior}{P(\hat{r}|\zeta, \eta, \hbar)} \propto \overset{likelihood}{P(\zeta|\hat{r}, \eta)} \overset{prior}{P(\hat{r}|\hbar)}$, where $\hbar$ denotes hyperparameters for the prior distribution.

We follow the CRP metaphor where the table assignment $t_{c_m} = t_k$ indicates that an observed path ζ_{c_m} belongs to the table t_k. This indicates that the path is generated by the agent with reward function $\hat{r}_{t_k}$. Let $K \rightarrow \infty$, given a set of observed agent paths represented as customers entering a restaurant $\zeta = \{\zeta_{c_m}\}_{c_m=1}^{M}$ and a set of latent parameters $\{\theta_{c_m}\}_{c_m=1}^{M}$, the PUR-IRL algorithm constructs a generative model in which the table $t_{c_m} = t_k$ assigned to a path ζ_{c_m} is defined by the latent parameter θ_{c_m} drawn according to the mixture model $\theta_{c_m}|G \sim G$, where $G|\alpha, G_0 \sim CRP(\alpha, d, G_0)$. After the reward function $\hat{r}_{t_k}$ is drawn from the prior $P(\hat{r}) = \prod_{f=1}^{F} P(r_f)$, the observed path ζ_{c_m} is drawn from the likelihood $P(\zeta_{c_m}|\hat{r}_{c_m}, \eta)$ given by (1). The reward function can be defined as follows:

$$\hat{r} = \mathbf{w} \cdot \boldsymbol{\gamma} \tag{3}$$

$$R(s, a) = \sum_{f}^{F} w_f \cdot \gamma_f(s, a), \tag{4}$$

where $\mathbf{w} : F \rightarrow [0, 1]$ represents the weight vector sampled from the prior and $\boldsymbol{\gamma} : S \times A \times F \rightarrow \{0, 1\}$ denotes a binary feature function indicating which reward features are relevant for each state-action pair. The joint posterior of the

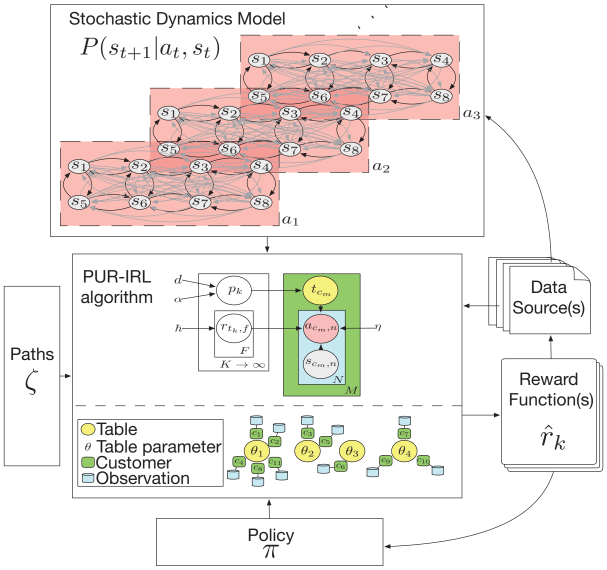

Figure 2: PUR-IRL framework

restaurant's seating arrangement $\vec{S} = \{t_{c_m}\}_{m=1}^{M}$ and the set of reward functions $\{\hat{r}_{t_k}\}_{k=1}^{K}$ is defined as follows:

$$P(\vec{S}, \{\hat{r}_{t_k}\}_{k=1}^{K}|\zeta, \eta, \alpha, d, \hbar) = \tag{5}$$

$$P(\vec{S}|\alpha, d) \prod_{k=1}^{K} P(\hat{r}_{t_k}|\zeta_{\vec{S}(t_k)}, \eta, \hbar), \tag{6}$$

where $\zeta_{\vec{S}(t_k)} = \{\zeta_{t_{c_m}} | t_{c_m} = t_k \text{ for } t_{c_m} = t_{c_1}, ..., t_{c_M}\}$.

Inference Procedure. To infer the latent reward functions from a set of paths, we approximate the full posterior joint distribution over the set of random variables via Bayesian inference with Metropolis-Hastings MCMC (MH-MCMC) sampling (Hastings 1970). MH-MCMC makes use of the full joint density function and (independent) proposal distributions for each variable of interest to simulate samples from a probability distribution. Given K unique table index values $\{t_1, ..., t_K\}$ in the restaurant, we can define the posterior distribution for table t_{c_m} as:

$$P(t_{c_m}|\vec{S}_{\backslash c_m}, \{\hat{r}_{t_k}\}_{k=1}^{K}, \zeta, \eta, \alpha, d) \tag{7}$$

$$\propto \overset{likelihood}{P(\zeta_{c_m}|\hat{r}_{t_{c_m}}, \eta)} \overset{prior}{P(t_{c_m}|\vec{S}_{\backslash c_m}, \alpha, d)}$$

$$P(t_{c_m}|\vec{S}_{\backslash c_m}, \alpha, d) \propto \begin{cases} \frac{count_{\backslash c_m, t_{c_j}} - d}{M+\alpha} \text{ if } t_{c_m} = t_{c_j} \\ \frac{\alpha + Kd}{M+\alpha} \text{ if } t_{c_m} \neq t_{c_j}, \end{cases} \tag{8}$$

where $\vec{S}_{\backslash c_m} = \{t_{c_i} | c_i \neq c_m \text{ for } c_i = c_1, c_2, ..., c_M\}$, and $count_{\backslash c_m, t_{c_j}}$ is the number of paths, excluding the current path, assigned to table t_{c_j}. Furthermore, if the sampled table t_{c_m} for path ζ_{c_m} is assigned to a new table, a new reward function $\hat{r}_{t_k}$ can be drawn from the distribution:

$$P(\hat{r}_{t_k}|\vec{S}, \hat{r}_{\backslash t_k}, \zeta, \eta, \hbar) \propto \overset{likelihood}{P(\zeta_{\vec{S}(t_k)}|\hat{r}_{t_k}, \eta)} \overset{prior}{P(\hat{r}_{t_k}|\hbar)} \tag{9}$$

Algorithm 1 PUR-IRL

```
Initialize $S^0, A^0, P^0$
for each pop-up $i \leq pIters$ do
  Update $MDP \backslash R = (S^{i-1}, A^{i-1}, P^{i-1})$
  Initialize restaurant seating arrangement $\vec{S}$ and set of reward functions $\{\hat{r}_{t_k}\}_{k=1}^{K}$ for all of its tables $\{t_k\}_{k=1}^{K}$
  for each crp $j \leq cIters$ do
    for each customer $c_m \leq totalDemos$ do
      $t^*_{c_m} \sim P(t|\vec{S}_{\backslash c_m}, \alpha, d)$
      if $t^*_{c_m} \notin \vec{S}_{\backslash c_m}$ then
        $\hat{r}_{t^*_{c_m}} \sim P(\hat{r}|\hbar)$
      end if
      $t_{c_m} = t^*_{c_m}$ and $\hat{r}_{t_{c_m}} = \hat{r}_{t^*_{c_m}}$
      w.p $min\left\{1, \frac{P(\zeta_{c_m}|\hat{r}_{t^*_{c_m}}, \eta)}{P(\zeta_{c_m}|\hat{r}_{t_{c_m}}, \eta)}\right.$
    end for
    for each table $t_k \leq totalTables$ do
      $\epsilon \sim \mathcal{N}(0, 1)$
      $\hat{r}^*_{t_k} = \hat{r}_{t_k} + \frac{1}{2}\tau^2 \nabla log f(\hat{r}_{t_k}) + \tau\epsilon$
      $\hat{r}_{t_k} = \hat{r}^*_{t_k}$ w.p $min\left\{1, \frac{f(\hat{r}^*_{t_k}) \times g(\hat{r}^*_{t_k}, \hat{r}_{t_k})}{f(\hat{r}_{t_k}) \times g(\hat{r}_{t_k}, \hat{r}^*_{t_k})}\right.$
    end for
  end for
  Use features in $\{\hat{r}_{t_k}\}_{k=1}^{K}$ with maximum posterior-probability to refine $S^i, A^i, P^i$
end for
```

Following random initialization of restaurant seating arrangement and its corresponding reward functions, the PUR-IRL algorithm begins an iterative procedure in which it perform two update operations. In the first update operation, the seating arrangement $\vec{S}$ is updated by sampling a new table index $t^*_{c_m}$ for each customer c_m according to Equation (7). If this new table index does not exist in the current seating arrangement $\vec{S}_{\backslash c_m}$, a new reward function is drawn from the reward prior. In the second update operation, each reward function $\hat{r}_{t_k}$ is updated by using a Langevin gradient update rule (Choi and Kim 2012). Following the CRP, the set of features associated with reward functions with the highest posterior probability are used for updating the S, A, P in the next pop-up restaurant iteration. Using the inferred optimal policy and reward function weights to prioritize which states and actions need to be updated, additional data sources (i.e. external functional, clinical databases, etc.) can be incrementally integrated into the MDP architecture in a tractable manner.

4 Real-World Use Case

4.1 The Colorectal Cancer Reward Function

Prior IRL methods have been applied in settings where reward function approximators with well-defined MDP's may suffice (i.e. GridWorld, route planning) (Choi and Kim 2012; Ziebart et al. 2008). We have designed an IRL experiment that is significantly more challenging—namely, the reconstruction of the evolutionary trajectories of CRC directly from tumor WGS data. With advances in bioinformatics and genomic sequencing, significant progress has been made in our understanding of CRC as a disease with multiple molecular subtypes (Guinney et al. 2015), distinct genetic trajectories for progression (Dickinson et al. 2015), and distinct modes of evolution (Sottoriva et al. 2015; Kim et al. 2018). This knowledge has been fueled by large cohorts with strong statistical analyses (Guinney et al. 2015), subclonal phylogenies (Deshwar et al. 2015; Sottoriva et al. 2015), and expert curation (Dickinson et al. 2015; Kim et al. 2018). However, this knowledge is largely incomplete due to the massive computational complexity of trying to acquire, analyze and model molecular properties across multiple scales. This poses a combinatorial problem that is currently heavily reliant on manual expertise.

Embracing Uncertainty in the MDP Structure of Cancer. Defining states and actions for IRL can be treated similarly to problems of feature representation, feature selection and feature engineering in unsupervised and supervised learning. For cancer data, we utilize the *Generalized Latent Feature Model* (GLFM) (Valera et al. 2017). Here, a *state* is encoded by a binary sparse code that indicates the presence/absence of latent features, inferred via GLFM, on the nucleotide, gene, and functional pathway level. An *action* then represents a stochastic event such as a somatic mutation in a specific gene. In addition to generating binary codes which provide more interpretable latent profiles of states and actions in the biological domain, the GLFM's use of a stochastic prior over infinite latent feature models allows model complexity to be adjusted on the basis of observations that will increase in volume and dimensionality as new data sources are incorporated in the PUR-IRL MDP.

Our initial MDP structure consists of 1084 actions and 144 states. An action corresponds to an event occurring at one of 1084 known 'driver' genes of CRC aggregated from two public datasets (Bamford et al. 2004; Tomczak, Czerwińska, and Wiznerowicz 2015). For example, action a_0^{AATK} corresponds to a mutation event occurring within any region of the *AATK* gene. The state space consists of 144 possible states composed of 12 latent features that were inferred via the GLFM algorithm. A state is an abstract representation that encodes features that are present internally or externally to a cancer cell (agent). The GLFM algorithm was used to infer these latent features from the list of alterations attributed to each inferred subclone. In this experiment, each state is represented by a 12-dimensional binary vector indicating the presence/absence of the 12 latent features inferred via the GLFM algorithm. Each latent feature reflects a unique frequency distribution of alterations to genes in 14 signaling pathways associated with CRC (Notch, Hedgehog, WNT, Chromatin Modification, Transcription, DNA damage, TGFβ, MAPK, STAT-JAK, PI3K-AKT, RAS, Cell-cycle, Apoptosis, Mismatch Repair). To infer a set of latent features, each subclone must be converted into a 14-dimensional vector indicating the count of alterations attributed to each signaling pathway. This set of 14-dimensional vectors serves as input to the GLFM algorithm.

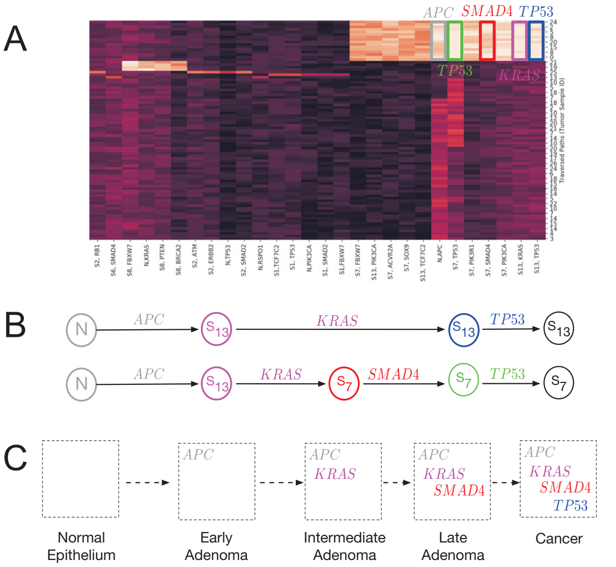

Figure 3: Summary results of PUR-IRL run on 27 CRC patient tumors. A) Heatmap of state/action pairs with highest reward values; B) Optimal paths derived from reward function with highest posterior-probability; C) Schematic presentation of the correlation between genetic changes and stages of colon cancer progression known as the "Vogelgram"

Embracing Uncertainty in Tumor Subclone Expert-Demonstrations. WGS data was used to infer the subclonal composition of a tumor using a slightly modified PhyloWGS algorithm (Deshwar et al. 2015) for efficiently identifying multiple possible unique phylogenetic trees. In a preliminary run of this experiment, 215,000 traversed paths derived from phylogenetic trees generated from a subset of (*N=27*) tumor samples were provided as *expert-demonstrations* to the PUR-IRL algorithm; with each path describing an ordered list of subclones within a given tumor sample and represented by a corresponding sequence of state-action pairs. The PUR-IRL model was run with 6 'pop-up' updates between every 100 CRP iterations.

Figure 3 summarizes the inferred reward function with highest posterior probability from this preliminary run. Figure 3.A shows a subset of the inferred reward function across the 27 tumor dataset. The optimal policy generated over this reward function consists of the state-action pairs *N-APC*, S_{13}-*KRAS*, S_7-*SMAD4*, highlighted in grey, pink, and red, respectively. The actions in these pairs correspond to genetic changes that are known to characterize CRC progression (Dienstmann et al. 2017) as summarized in Figure 3.C. We compare this to the most likely paths drawn in Figure 3.B that were obtained by simulating a MDP with our new reward function. Despite uncertainties in how our data was generated, we were able to recapitulate an optimal path, or evolutionary trajectory, with biologically relevant conclusions which match the literature derived model of CRC progression (Fearon and Vogelstein 1990). This demonstrates the PUR-IRL model's ability to identify singular genetic

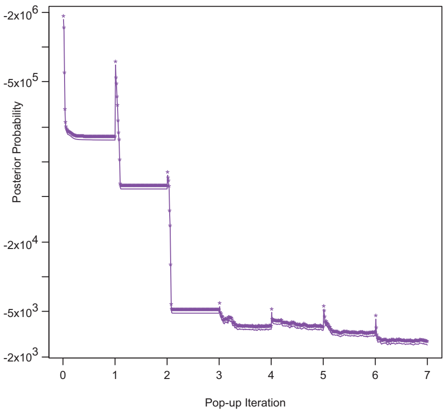

Figure 4: Posterior Probability of Inferred Reward Functions during PUR-IRL Iterations

changes that are often not the most frequent, but are nonetheless critical for CRC progression. In Figure 4, we analyze the posterior probability of the inferred reward functions following each of the 6'pop-up'updates and 100 intermediate CRP iterations. The results of this analysis demonstrate that incremental integration of new information by PUR-IRL provides a tractable methodology for improving the reward functions without the use of a large initial state or action-space while still allowing for the exploration of new features.

5 PUR-IRL Performance

In order to demonstrate the PUR-IRL algorithm's utility and accuracy, we ran PUR-IRL on data generated by multiple experts and sampled under uniform and non-uniform sampling conditions. Specifically, we performed three sets of experiments (130 total) that evaluated the performance of PUR-IRL on the *GridWorld* problem (Abbeel and Ng 2004). In each experiment, we evaluate the IRL model under different discount hyperparameter values $\{0.0, 0.3, 0.7, 1.0\}$, where PUR-IRL with $d = 0$ reduces to the DPM-IRL. The GridWorld is a simple deterministic world that is often used to illustrate the basic concepts of Q-learning. It allows us to evaluate our approach under different scenarios in which the inference of latent reward functions can be validated using simulated ground-truths.

We consider an 8×8 GridWorld, where each of the 64 cells corresponds to a state representing to the location of the agent on the grid. The agent can execute one of four possible actions in order to move north, south, east or west. The execution of an action has a 20% chance of failing and resulting in a random move to one of the adjacent states. The grid is divided into non-overlapping subregions of 2×2 cells. A small number of the 16 subregions has a positive reward associated with them. For each $i = 1, ..., 16$, there is one feature $\gamma_i(s)$ indicating whether the state s is in subregion i. Thus, the rewards may be written $\hat{r} = \mathbf{w}^{\mathbf{T}} \cdot \gamma$. The weights $\mathbf{w}$ can be sampled from a prior distribution and the initial state drawn from a uniform distribution on the states.

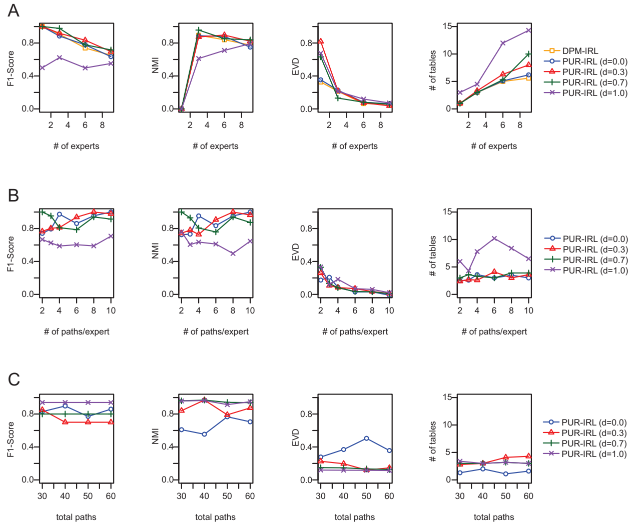

Figure 5: GridWorld Simulation Results

In the first set of experiments, the GridWorld conditions from (Choi and Kim 2012)—3 expert demonstrations generated per expert—were amended to explicitly model 4 scenarios in which the number of experts is greater than or equal to one. In each scenario, we randomly sample the weights for the G ground-truth reward functions (experts), $\{r^g_1, r^g_2, r^g_3 ... r^g_G\}$ from a Gaussian prior and evaluate IRL performance under uniform sampling conditions (i.e. each expert generates the same number of paths). This experiment was repeated 10 times for each scenario. The results of this experiment (Fig. 5A) demonstrate that PUR-IRL and other IRL methods that use Bayesian nonparametric priors (DPM-IRL) can recapitulate the ground-truth reward function(s) from data that follows the single expert assumption in addition to scenarios where the true number of data-generating experts is unknown. As an additional sanity check, we were able to verify that performance of PUR-IRL ($d = 0$) model, which reduces the underlying PYP to a DP, mimics that of DPM-IRL and the results found in (Choi and Kim 2012). In the second experiment set, we randomly sample the weights for 3 ground-truth reward functions, under 5 uniform sampling conditions of increasing dataset size. Results from this experiment (Fig. 5B) show that with an increase in the number of paths, PUR-IRL model performance improves in terms of the number of tables (inferred reward functions), normalized mutual information, F1-scores and the expected value difference (EVD) between the ground-truth reward functions and the learned reward functions. Unsurprisingly, the outlier model (PUR-IRL with $d = 1.0$) fails to improve in performance or accurately recapitulate the true number of reward functions (tables) due its heavy bias for fat-tail distributions. In the final experiment set we sought to evaluate IRL performance under 4 non-uniform sampling conditions which closely resemble those found in real-world data (i.e the total set of paths is distributed across 3 experts according to a power-law distribution). In Figure 5C, we can see that the addition of the discount hyperparameter within the PUR-IRL model allows us to control how well the final model fits with input dataset and thus allowing us to exceed performance when $d = 0$. Although the GridWorld MDP does not encapsulate many of the complexities that we address with our framework (i.e. ability to infer the number and identity of biologically relevant states and actions from high-dimensional data), it nevertheless demonstrates that PUR-IRL can accurately infer optimal policies and latent reward functions given a set of expert demonstrations under various data scenarios likely to be found in real-world applications.

6 Discussion

In this paper, we explore the use of IRL as a viable approach for distilling knowledge about a complex decision-making process from ambiguous and problematic tumor data. To do so, we introduce and evaluate the PUR-IRL algorithm and its ability to use expert demonstrations of cancer evolution from patient tumor WGS data. We demonstrate that by formalizing cancer behavior as a MDP, the state-action pairs highlighted by the inferred reward function and optimal policy can be used to reach interpretable biological conclusions. Furthermore, we were able to show that the incremental integration of new information through iterative MDP structural updates allows for improvements in the posterior probability of our reward function in an adaptive manner that is amenable to new input data. Finally, we were able to recapitulate ground truth reward functions from simulated expert demonstrations using GridWorld, demonstrating PUR-IRL's ability to infer reward functions despite uncertainties about the source and structure of our input data.

Cancer evolution remains one of the most important and intractable problems in the natural sciences—one that sits at the intersection of myriad biological disciplines ranging from molecular and developmental biology to stochastic processes and population dynamics. Each of these research areas and their associated datasets sheds light on important pieces of the biological puzzle, but data integration and knowledge synthesis requires a unified framework. It is our hope that the development of unreasonably effective algorithms such as PUR-IRL will advance our understanding of the complex structure and relationships inherent in cancer data. Furthermore, we explicitly choose to do so using an IRL approach that allows us to quantify the influence of intrinsic and extrinsic factors on cancer progression while also accounting for uncertainty. This provides a mechanistic time-ordered event history of cancer, granting us a window into causality using data derived from observational studies of human tumors.

Acknowledgments. We would first like to thank the patients who volunteered for the original study. We also thank the many other individuals who made this work possible including members of the Mayo Clinic Microbiome Laboratory, study coordinators, students, colorectal surgeons, program directors, and pathology assistants. We also specially acknowledge the assistance of Gregory Dougherty during software development and Dr. Lisa Boardman for her clinical insight. We would like to thank the National Cancer Institute (R01CA179243), the DeWitt and Curtiss families, the Center for Individualized Medicine at Mayo Clinic for their generous support of this work. Last but not least, Mathieu Wiepert, Marie Kostecki, and Shamailah Lando for keeping the high-performance computing clusters running through our exhaustive calculations.

References

Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1. ACM.

Bamford, S.; Dawson, E.; Forbes, S.; Clements, J.; Pettett, R.; Dogan, A.; Flanagan, A.; Teague, J.; Futreal, P. A.; Stratton, M. R.; et al. 2004. The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer* 91(2):355.

Burrell, R. A.; McGranahan, N.; Bartek, J.; and Swanton, C. 2013. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501(7467):338.

Choi, J., and Kim, K.-E. 2012. Nonparametric bayesian inverse reinforcement learning for multiple reward functions. In *Advances in Neural Information Processing Systems*, 305–313.

Cingolani, P.; Platts, A.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.; Lu, X.; and Ruden, D. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2):80–92.

DePristo, M. A.; Banks, E.; Poplin, R.; Garimella, K. V.; Maguire, J. R.; Hartl, C.; Philippakis, A. A.; Del Angel, G.; Rivas, M. A.; Hanna, M.; et al. 2011. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics* 43(5):491.

Deshwar, A. G.; Vembu, S.; Yung, C. K.; Jang, G. H.; Stein, L.; and Morris, Q. 2015. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology* 16(1):35.

Dickinson, B. T.; Kisiel, J.; Ahlquist, D. A.; and Grady, W. M. 2015. Molecular markers for colorectal cancer screening. *Gut* 64(9):1485–1494.

Dienstmann, R.; Vermeulen, L.; Guinney, J.; Kopetz, S.; Tejpar, S.; and Tabernero, J. 2017. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer* 17(2):79.

El-Kebir, M.; Oesper, L.; Acheson-Field, H.; and Raphael, B. J. 2015. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 31(12):i62–i70.

Fearon, E. R., and Vogelstein, B. 1990. A genetic model for colorectal tumorigenesis. *cell* 61(5):759–767.

Guinney, J.; Dienstmann, R.; Wang, X.; De Reynies, A.; Schlicker, A.; Soneson, C.; Marisa, L.; Roepman, P.; Nyamundanda, G.; Angelino, P.; et al. 2015. The consensus molecular subtypes of colorectal cancer. *Nature medicine* 21(11):1350.

Ha, G.; Roth, A.; Khattra, J.; Ho, J.; Yap, D.; Prentice, L. M.; Melnyk, N.; McPherson, A.; Bashashati, A.; Laks, E.; et al. 2014. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research* 24(11):1881–1893.

Hale, V. L.; Jeraldo, P.; Chen, J.; Mundy, M.; Yao, J.; Priya, S.; Keeney, G.; Lyke, K.; Ridlon, J.; White, B. A.; et al. 2018a. Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers. *Genome medicine* 10(1):78.

Hale, V. L.; Jeraldo, P.; Mundy, M.; Yao, J.; Keeney, G.; Scott, N.; Cheek, E. H.; Davidson, J.; Green, M.; Martinez, C.; et al. 2018b. Synthesis of multi-omic data and community metabolic models reveals insights into the role of hydrogen sulfide in colon cancer. *Methods* 149:59–68.

Halevy, A.; Norvig, P.; and Pereira, F. 2009. The unreasonable effectiveness of data.

Hastings, W. K. 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57(1):97–109.

Johnson, N., and Kotz, S. 1977. *Urn models and their application: an approach to modern discrete probability theory*. Wiley Series in Probability and Statistics: Applied Probability and Statistics Section. Wiley.

Kim, M.; Druliner, B. R.; Vasmatzis, N.; Bae, T.; Chia, N.; Abyzov, A.; and Boardman, L. A. 2018. Inferring modes of evolution from colorectal cancer with residual polyp of origin. *Oncotarget* 9(6):6780.

Koza, J. R. 1992. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics* 25(14):1754–1760.

Li, Y.; Mahjoubfar, A.; Chen, C. L.; Niazi, K. R.; Pei, L.; and Jalali, B. 2019. Deep cytometry: Deep learning with real-time inference in cell sorting and flow cytometry. *Scientific reports* 9(1):11088.

Ng, A. Y., and Russell, S. 2000. Algorithms for inverse reinforcement learning. In *in Proc. 17th International Conf. on Machine Learning*. Citeseer.

Schwartz, R., and Schäffer, A. A. 2017. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* 18(4):213.

Siegel, R. L.; Miller, K. D.; and Jemal, A. 2019. Cancer statistics, 2019. *CA: a cancer journal for clinicians* 69(1):7–34.

Sottoriva, A.; Kang, H.; Ma, Z.; Graham, T. A.; Salomon, M. P.; Zhao, J.; Marjoram, P.; Siegmund, K.; Press, M. F.; Shibata, D.; et al. 2015. A big bang model of human colorectal tumor growth. *Nature genetics* 47(3):209.

Tomczak, K.; Czerwińska, P.; and Wiznerowicz, M. 2015. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology* 19(1A):A68.

Valastyan, S., and Weinberg, R. A. 2011. Tumor metastasis: molecular insights and evolving paradigms. *Cell* 147(2):275–292.

Valera, I.; Pradier, M. F.; Lomeli, M.; and Ghahramani, Z. 2017. General latent feature models for heterogeneous datasets. *arXiv preprint arXiv:1706.03779*.

Wang, K., and Wan, X. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, 4446–4452.

Wigner, E. 1960. The unreasonable effectiveness of mathematics in the natural sciences. *Comm. Pure and Applied Mathematics* 13:001–14.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning.